COURSE PACK

Tuberculosis Diagnostic Research: Beyond the Basics

An intensive course on TB diagnostic research methods - from basics to advanced techniques

December 13 - 15, 2010

Tuberculosis Research Centre, Chennai, India

ICMR CENTENARY CELEBRATIONS

Indian Council of Medical Research

V.Ramalingaswami Bhawan, Ansari Nagar, New Delhi-110 029

TDR For research on diseases of poverty

UNICEF • UNDP • World Bank • WHO

This course is organized with funding support from TDR, and additional support from European Commission, EDCTP, McGill University, Stop TB Partnership's New Diagnostics Working Group & Foundation for Innovative New Diagnostics

Hosted by

Tuberculosis Research Centre
Indian Council of Medical Research
WHO Collaborating Centre for Tuberculosis Research and Training

ICMR CENTENARY CELEBRATIONS

Indian Council of Medical Research

V.Ramalingaswami Bhawan, Ansari Nagar, New Delhi-110 029

Tuberculosis Diagnostic Research: Beyond the Basics

December 13 - 15, 2010, Tuberculosis Research Centre, Chennai, India

COURSE SCHEDULE

Monday, Dec 13, 2010

Time	Lecture	Faculty
8.30 AM	Welcome and Introductions	A Thomas, MS Jawahar, all
9.00 AM	The global value chain (blueprint) for TB diagnostics and current pipeline of diagnostics	M Pai
9.45 AM	New WHO policy on Xpert MTB/RIF New WHO policy on serological assays	CN Paramasivan K Steingart
10.30 AM	Coffee	
10.45 AM	Overview of diagnostic research and types of diagnostic study designs	M Pai
11.15 AM	Landscape of TB diagnostic research	M Pai
12.30 PM	Lunch	
1.30 PM	The diagnostic test accuracy design	M Pai
2.30 PM	Bias in diagnostic research and sources of variation	K Steingart
3.30 PM	Coffee	
4.00 - 5.30 PM	Optimism bias in TB diagnostic research & critical appraisal	M Pai

Tuesday, Dec 14, 2010

Time	Lecture	Faculty
8.30 AM	Setting reference standards in TB diagnostic studies: • Microscopy evaluations • Rapid culture methods • Molecular assays • Serological assays • IGRAs and LTBI tests • Extrapulmonary TB	 CN Paramasivan CN Paramasivan T Shinnick S Laal M Pai J Peter
10.30 AM	Coffee	
10.45 PM	Lab accreditation and QA in India Ensuring quality in diagnostic trials: • Microscopy studies • Culture studies • Molecular assay studies • Immunodiagnostic studies	J Kenneth CN Paramasivan CN Paramasivan T Shinnick S Laal

12.00 Noon	Introduction to TB biomarker validation	S Parida
12.30 PM	Lunch	
1.30 PM	Beyond test accuracy - 1: impact of testing on diagnostic thinking and clinical decision making	M Pai
2.00 PM	Beyond test accuracy - 2: incremental value and multivariable methods	M Pai
3.30 PM	Coffee	
4.00 PM	Beyond test accuracy - 3: randomized designs for clinical impact	M Pai J Peter
4.45 - 5.30 PM	Evaluation of TB diagnostics in children and HIV-infected populations: challenges and potential solutions	S Swaminathan

Wednesday, Dec 15, 2010

Time	Lecture	Faculty
8.30 AM	Beyond test accuracy - 4: cost and cost-effectiveness	H Sohn
9.30 AM	Meta-analysis of diagnostic research	K Steingart
10.30 AM	Coffee	
11.00 AM	Beyond test accuracy - 5: Guideline and policy development using the GRADE approach	K Steingart
11.45 AM	Converting ideas into commercially viable products	N Sriram
12.30 PM	Lunch	
1.30 PM	Landscape of TB diagnostics in India and barriers to innovation	M Pai
2.00 PM	Panel discussion on diagnostic innovations in India	P Small S Swaminathan M Pai (moderator) + course participants
3.30 PM - 5.00 PM	**ICMR Centenary Celebration Special Lecture**: *TB control in India: what is the critical path after DOTS scale-up?* **Peter M Small, MD** Senior Program Officer, TB Global Health Program Bill & Melinda Gates Foundation, USA	Co-chairs: PR Narayanan & S Swaminathan

This course is organized with funding support from TDR, and additional support from European Commission, EDCTP, McGill University, Stop TB Partnership's New Diagnostics Working Group & Foundation for Innovative New Diagnostics

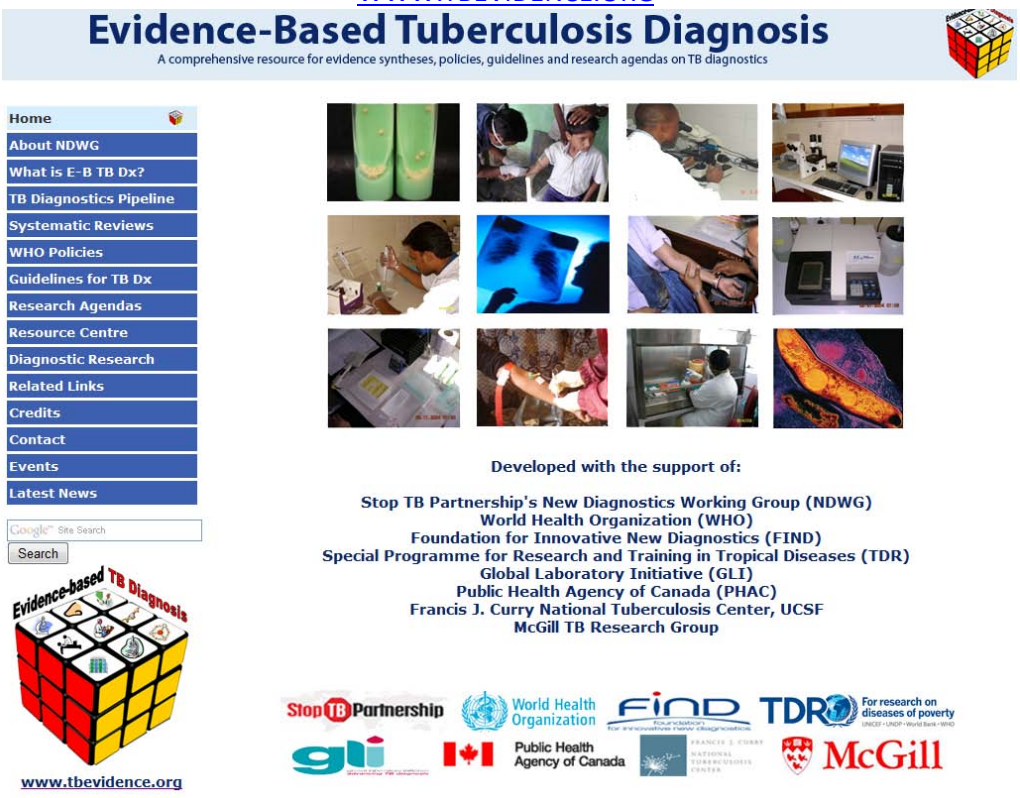
Course faculty

Karen R Steingart, MD Physician Consultant Curry International Tuberculosis Center University of California, San Francisco San Francisco, USA Email: karenst@uw.edu	**Madhukar Pai, MD, PhD** Assistant Professor, McGill University Co-chair, Stop TB Partnership's New Diagnostics Working Group Dept of Epidemiology & Biostatistics Montreal, Canada Email: madhukar.pai@mcgill.ca
Jonathan Peter, MD Clinical trials co-ordinator/research fellow UCT Lung Infection and Immunity Unit and Lung Institute, Department of Medicine University of Cape Town South Africa Email: jonnyp@mweb.co.za	**Soumya Swaminathan, MD** Coordinator (Research for Neglected Priorities) UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases World Health Organization Geneva, Switzerland Email: swaminathans@who.int
Suman Laal, PhD Associate Professor of Pathology & Microbiology NYU Langone Medical Center New York, USA Email: Suman.Laal@nyumc.org	**C.N. Paramasivan, Ph.D., D.Sc.** Head of TB Laboratory Support Foundation for Innovative New Diagnostics Geneva, Switzerland E-mail: cn.paramasivan@finddiagnostics.org
Thomas M. Shinnick, Ph.D. Associate Director for Global Laboratory Activities Division of Tuberculosis Elimination Centers for Disease Control, Atlanta USA Email: TMS1@CDC.GOV	**Shreemanta K Parida, MD, PhD** TB research specialist Berlin Germany Email: shreemanta.parida@googlemail.com
N Sriram Director - Tulip Group Verna, Goa India E-mail: orchid@tulipgroup.com	**John Kenneth, MD** Head, Division of Infectious Diseases St. Johns Research Institute St. Johns National Academy of Health Sciences Bangalore, India Email: johnkennet@gmail.com
Peter M Small Senior Program Officer, TB Global Health Program Bill & Melinda Gates Foundation Seattle, USA	**Hojoon Sohn, MPH** Doctoral candidate McGill University Dept of Epidemiology & Biostatistics Montreal, Canada Email: dhjsohn@gmail.com

यक्ष्मा अनुसंधान केन्द्र
YAKSHMA ANUSANDHANA KENDRA
TUBERCULOSIS RESEARCH CENTRE
(Indian Council of Medical Research)
Mayor V.R. Ramanathan Road
Chetput, Chennai 600 031

KEY WEBSITE RESOURCES

WWW.TBEVIDENCE.ORG



WWW.TEACHEPI.ORG

OPEN ACCESS Freely available online

PLoS MEDICINE

Research in Translation

Evidence-Based Tuberculosis Diagnosis

Madhukar Pai*, Andrew Ramsay, Richard O'Brien

There is great excitement in the tuberculosis (TB) scientific community over the introduction of new tools into TB control activities. The development of new tools is an important component of the Global Plan to Stop TB and the World Health Organization's new global Stop TB Strategy [1,2]. Anticipating the introduction of new tools, the Stop TB Partnership has established a Retooling Task Force to develop a framework for engaging policy makers to foster accelerated adoption and implementation of new tools into TB control programs [3].

While new tools offer great promise in clinical medicine and in public health, limited resources and the movement toward evidence-based guidelines and policies require careful validation of new tools prior to their introduction for routine use. The world spends an estimated US$1 billion per year on diagnostics for TB [4]. It is important to ensure that such expenditure is backed by strong evidence.

Ideally, clinical and policy decisions must be guided by the totality of evidence on a given topic. This is particularly relevant for TB, where concerns have been raised about the lack of emphasis on evidence of effectiveness in some of the existing TB guidelines and policies [5]. These concerns are being taken seriously [6,7], and the outcome should be evident in upcoming TB guidelines and policies. In fact, the World Health Organization (WHO) recently announced its approach for developing new policies on TB in a document entitled "Moving Research Findings into New WHO Policies" [7]. According to this document, in order to consider a global policy change, WHO must have solid evidence, including clinical trials or field evaluations in high TB prevalence settings. The steps involved in the policy process include a comprehensive review of the evidence, as well as expert opinion and judgment (Box 1).

High-quality evidence on TB diagnostics is critical for the development of evidence-based policies on TB diagnosis, and, ultimately, for effective control of the global TB epidemic. While primary diagnostic trials are needed to generate data on test accuracy and operational performance, systematic reviews provide the best synthesis of current evidence on any given diagnostic test [8]. Although a large number of trials on TB diagnostics have been published, surprisingly, no systematic reviews were published until recently. In the past few years, at least 30 systematic reviews and meta-analyses have been published on various TB tests [9–38]. These reviews have synthesized the results of more than 1,000 primary studies, providing valuable insights into the diagnostic accuracy of various tests (Table 1, Box 2).

Implications for Clinical and Laboratory Practice

For clinicians, systematic reviews provide several useful insights for diagnosis of latent TB infection, active TB disease, and drug resistance.

For diagnosis of latent TB, clinicians have used the tuberculin skin test (TST) for decades. Recently, interferon-gamma release assays (IGRAs) have emerged as attractive alternatives. While the TST is known to have poor specificity in populations vaccinated with bacille Calmette-Guérin (BCG) [34], meta-analyses have shown that IGRAs have much higher specificity for TB infection than the TST, and IGRA specificity is unaffected by BCG vaccination [21,26,37]. However, another meta-analysis showed that BCG vaccination received in infancy has a minimal effect on the TST, whereas BCG received after infancy produces more frequent, more persistent, and larger TST reactions [35]. Thus, the TST might retain high specificity in some populations, whereas it may perform poorly in others. IGRAs are particularly attractive in the latter setting. However, meta-analyses on IGRAs have highlighted the lack of evidence on the predictive ability of these assays in identifying those individuals with TB infection who are at highest risk for progressing to active disease. Several cohort studies are ongoing (reviewed elsewhere [39]), and these should provide useful evidence on this unresolved issue.

For active TB, serological tests have been attempted for decades. Two meta-analyses have convincingly shown that existing commercial antibody-based tests have poor accuracy and limited clinical utility [29,30]. Despite this evidence, dozens of commercial serological tests continue to be marketed, mostly in private sectors of countries that lack diagnostic regulatory bodies [4].

Nucleic acid amplification tests (NAATs) were considered to be a major breakthrough in TB diagnosis when they were first introduced. A series of meta-analyses have shown that NAATs have high specificity and positive predictive value, but modest and highly variable sensitivity, especially

Research in Translation discusses health interventions in the context of translation from basic to clinical research, or from clinical evidence to practice.

Citation: Pai M, Ramsay A, O'Brien R (2008) Evidence-based tuberculosis diagnosis. PLoS Med 5(7): e156. doi:10.1371/journal.pmed.0050156

Abbreviations: ADA, adenosine deaminase; BCG, bacille Calmette-Guérin; FIND, Foundation for Innovative New Diagnostics; IFN-γ, interferon-gamma; IGRA, interferon-gamma release assay; LED, light-emitting diode; MDR-TB, multidrug-resistant tuberculosis; NAAT, nucleic acid amplification test; STAG-TB, Strategic and Technical Advisory Group for Tuberculosis; TB, tuberculosis; TDR, UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases; TST, tuberculin skin test; WHO, World Health Organization; XDR-TB, extensively drug-resistant tuberculosis

Madhukar Pai is with the Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada. Andrew Ramsay is with the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland. Richard O'Brien is with the Foundation for Innovative New Diagnostics, Geneva, Switzerland.

* To whom correspondence should be addressed. E-mail: madhukar.pai@mcgill.ca

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]

in smear-negative and extrapulmonary TB [9,11,14,18,23,24,28].

Conventional tests such as smears and cultures perform poorly in extrapulmonary TB. A series of reviews have shown that biomarkers such as adenosine deaminase (ADA) and interferon-gamma (IFN-γ) have excellent accuracy for tuberculous pleural effusion [12,13,15,17]. These biomarkers, especially ADA, are easy to measure and inexpensive. Despite this evidence, these tests appear to be underutilized [40].

For the diagnosis of multidrug-resistant TB (MDR-TB), available data suggest that phage-based assays do not perform well when directly applied to clinical specimens [25]. Line probe assays show great promise for rapid detection of rifampicin resistance in settings with high MDR-TB prevalence [22,38]. Simple tests such as colorimetric redox methods and nitrate reductase assays appear to perform very well, but require culture isolation [19,36]. More evidence is needed on rapid tests for drug resistance, especially since the Global XDR-TB Response Plan calls for wide-scale implementation of rapid methods to screen patients at risk of XDR-TB (extensively drug-resistant TB) and MDR-TB [41].

For laboratory practice, systematic reviews provide strong evidence that fluorescence microscopy is more sensitive than conventional light microscopy (with no significant loss in specificity) [31], that sputum processing methods (e.g., bleach or centrifugation) can be effective in increasing the yield of smear microscopy [32], and that liquid cultures are more rapid and sensitive than solid cultures [10].

Implications for Policies and Guidelines

In addition to informing evidence-based TB diagnosis, systematic reviews have been helpful in informing policy decisions. For example, a series of recent reviews has shown that smear microscopy can be optimized using at least three different approaches: chemical and physical processing for concentration of sputum, use of fluorescence microscopy instead of conventional light microscopy, and the examination of two (as compared to three) sputum specimens [20,31,32]. The findings of these reviews were incorporated into the International Standards for TB Care [42], and have informed policy guidance on the diagnosis of smear-negative TB in people living with HIV/AIDS [43].

The review on incremental yield of serial smears showed that the average incremental yield and/or increase in sensitivity of examining a third sputum specimen ranged between 2% and 5% [20]. This suggested that reducing the recommended number of specimens examined from three to two could potentially benefit TB control programs, and potentially increase case detection for several reasons [20]. Partly based on this evidence and expert opinion, WHO recently revised its policies on smear microscopy [44]. It now recommends that the number of specimens to be examined for screening of TB cases be reduced from three to two, in places where a well-functioning external quality assurance system exists, where the workload is very high, and where human resources are limited [44]. The revised WHO definition of a new sputum smear-positive pulmonary TB case is based on the presence of at least one acid fast bacillus in at least one sputum sample in countries with a well-functioning external quality assurance system [45].

These new policies have major implications for resource-poor settings with high TB prevalence where sputum microscopy is the main or

doi:10.1371/journal.pmed.0050156.g001

Figure 1. Low-Cost LED-Based Fluorescence Microscopy Being Evaluated at a TDR/WHO Trial Site in Abuja, Nigeria
Photographer: Andrew Ramsay (Courtesy of TDR, Geneva)

only diagnostic test available, and particularly where laboratory services are being overwhelmed with demand for smear microscopy. Omitting the third smear could potentially reduce costs and alleviate the workload of laboratories, particularly in countries with human resource crises. In these settings, laboratories performing smear microscopy often have to deal with anemia, malaria, and other diseases. Thus, the time saved from the inefficient examination of a third smear may be applied toward improving laboratory testing for other diseases [20]. The adoption of the revised case definition and a two-smear approach may create the opportunity to examine both smears during a patient's first presentation to a health facility, and thereby reduce the large numbers of patients known to drop out during the diagnostic process [46]. While these are reasonable assumptions, it is worth emphasizing that there is no hard evidence that the two-smear policy actually improves TB control in the real world. Such data will have to come from programmatic research at the country level and from data collected in routine public health program settings.

There is strong evidence that liquid cultures are more sensitive and rapid than solid media cultures [10]. Based on a review of available evidence and an expert consultation, WHO recently issued policy guidance on the use of liquid TB culture and drug susceptibility testing in low-resource settings [47]. The WHO policy recommends phased implementation of liquid culture systems as a part of a country-specific comprehensive plan for laboratory capacity strengthening that addresses issues such as biosafety, training, maintenance of infrastructure, and reporting of results [47]. These policies are expected to have an important impact in settings with high HIV prevalence [43] and in countries where MDR-TB is an increasing problem [41], helping to inform the needed global scale-up of culture and drug susceptibility testing capacity.

However, implementation of culture testing requires a well-functioning health care system, adequate laboratory infrastructure, and trained personnel. Therefore, emphasis must be placed on capacity building and health system and laboratory strengthening [43,48]. Recognizing this, the Stop TB Partnership, WHO, and partners have launched a Global Laboratory Initiative to facilitate laboratory policy guidance, technical assistance, quality management, resource mobilization, and advocacy. Again, as in the case of the two-smear strategy, it must be emphasized that there is no strong evidence that the WHO policy on liquid cultures actually improves TB control at the routine programmatic level. Field studies and cost-effectiveness data are needed to better understand the real world implications of this policy.

In June 2008, WHO announced a new policy statement, endorsing the use of line probe assays for rapid screening of patients at risk of MDR-TB (http://www.who.int/tb/en/). This policy statement was based in part on evidence summarized in systematic reviews [22,38], expert opinion, and results of field demonstration projects. The recommended use of line probe assays is currently limited to culture isolates and direct testing of smear-positive sputum specimens. Line probe assays are not recommended as a complete replacement for conventional culture and drug susceptibility testing. Culture is still required for smear-negative specimens, and conventional drug susceptibility testing is still necessary to confirm XDR-TB.

Following this new policy, WHO, UNITAID, the Stop TB Partnership, and the Foundation for Innovative New Diagnostics (FIND) announced a new initiative to improve the diagnosis and treatment of MDR-TB in resource-limited settings (http://www.who.int/tb/features_archive/mdrtb_rapid_tests/en/index.html). As part of this initiative, over the next few years, 16 countries will begin using rapid tests to diagnose MDR-TB, including line probe assays. The countries will receive specially priced tests through the Stop TB Partnership's Global Drug Facility, which provides countries with both drugs and diagnostic reagents.

Implications for Research and Development

Systematic reviews have been helpful in identifying key knowledge gaps and defining research agendas. For example, based on the smear microscopy reviews [20,31,32] and expert opinion, the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical

doi:10.1371/journal.pmed.0050156.g002

Figure 2. A Simplified NAAT Being Evaluated at a FIND Trial Site in India
Photographer: Ralf Linke (Courtesy of FIND, Geneva)

Diseases (TDR) recently launched a major research program aimed at the optimization of smear microscopy [49]. Large-scale field studies are ongoing in more than ten countries on issues such as optimum timing and composition of sputum specimen sets; use of lower-cost light-emitting diode (LED) fluorescence microscopy systems (Figure 1); sputum processing methods involving bleach digestion; and potential for reducing time to diagnosis and number of patient visits required by examining two specimens on the day that the patient first presents. The latter can be expected to reduce the considerable patient drop-out rates during diagnosis that are seen in many settings [46].

In parallel, FIND recently forged a partnership with Carl Zeiss MicroImaging (http://www.zeiss.com/micro/) to develop an inexpensive, robust LED-based microscope that will be extensively evaluated for routine use in high-burden countries [50].

Systematic reviews on existing commercial serological tests and NAATs have shown that these assays have not performed as well as expected [14,18,29,30]. A recent evaluation of 19 rapid commercial serological tests for TB using specimens from the TDR TB Specimen Bank confirmed the poor accuracy of existing serological tests for TB [51]. Such evidence has informed several initiatives to improve serological assays and NAATs. For example, FIND is supporting the development and evaluation of newer, improved NAATs (Figure 2) [52]. Several groups are working on methods to optimize serological assays, including the use of novel TB-specific antigens, the use of antigen combinations, and the development of point-of-care tests [52].

Systematic reviews on IGRAs have informed the development of guidelines and positions statements in many countries [53,54,55]. They have also facilitated the development of a comprehensive research agenda with a specific focus on the use of these assays in resource-limited settings [56].

Systematic reviews on TB diagnostics have revealed deficiencies in the quality of TB diagnostic trials. A recent analysis of systematic reviews showed that trials of TB diagnostics lack methodological rigor, and studies are often poorly reported [57]. Lack of methodological rigor in trials is a cause for concern, as it may prove to be a major hurdle for effective application of diagnostics in TB care and control. Biased results from poorly designed trials can lead to premature adoption of diagnostics that may have little or no benefit. The situation is exacerbated by the fact that most developing countries have poor regulatory mechanisms for licensing and post-marketing surveillance of diagnostics. For example, dozens of commercial serological tests are marketed in developing countries, despite lack of evidence on their utility [29,30,51].

It is clear that efforts are needed to improve both methodological quality and reporting of TB diagnostic trials [57,58]. TDR has developed guidelines for researchers on assessing the performance and operational characteristics of diagnostics for infectious diseases [59], and the STARD (Standards for Reporting

Box 2. Five Key Papers in the Field

Dinnes et al., 2007 [10]. The most comprehensive systematic review of several rapid diagnostic tests for the detection of TB, sponsored by the UK Health Technology Assessment Programme.

Mase et al., 2007 [20]. This review on incremental yield of serial smears showed that the average incremental yield and/or the increase in sensitivity of examining a third sputum specimen ranged between 2% and 5%. This evidence partly informed the new WHO policy on smear microscopy.

Menzies et al., 2007 [21]. This meta-analysis showed that IGRAs for TB infection have excellent specificity (higher than the conventional TST), and are unaffected by prior BCG vaccination. This review also highlighted the key unresolved questions regarding the use of these assays in clinical practice. An update to this meta-analysis was published recently (Pai et al., 2008 [37]).

Steingart et al., 2007 [30]. This meta-analysis showed that serological tests for TB produce highly inconsistent estimates of sensitivity and specificity, and none of the currently available commercial assays perform well enough to replace microscopy. Several initiatives are now ongoing to develop improved point-of-care immune-based rapid tests for TB.

Steingart et al., 2006 [31]. This systematic review reported strong evidence that fluorescence microscopy is more sensitive than conventional microscopy. Several initiatives are now ongoing to develop simple, low-cost fluorescence microscopy systems to optimize smear microscopy.

Table 1. Findings from Systematic Reviews on TB Diagnostic Tests

Diagnostic Test [References]	Number of Reviews	Disease/Site	Major Findings/Results of Systematic Reviews
Diagnosis of active TB			
Sputum smear microscopy [20,31,32]	3	Pulmonary TB	• Fluorescence microscopy is on average 10% more sensitive than conventional microscopy. Specificity of both fluorescence and conventional microscopy is similar. • Centrifugation and overnight sedimentation, preceded with any of several chemical methods (including bleach), is more sensitive than direct microscopy; specificity is unaffected by sputum processing methods. • When serial sputum specimens are examined, the mean incremental yield and/or increase in sensitivity from examination of 3rd sputum specimen ranges between 2% and 5%.
NAATs [9,10,11,14,18,23,24,28]	8	Pulmonary and extrapulmonary TB	• NAATs have high specificity and positive predictive value. NAATs, however, have relatively lower (and highly variable) sensitivity and negative predictive value for all forms of TB, especially in smear-negative and extrapulmonary disease. • In-house ("home-brew") NAATs produce highly inconsistent results as compared to commercial, standardized NAATs.
Commercial serological antibody detection tests [10,29,30]	3	Pulmonary and extrapulmonary TB	• Serological tests for both pulmonary and extrapulmonary TB produce highly inconsistent estimates of sensitivity and specificity; none of the assays perform well enough to replace microscopy.
ADA [12,13,17,27,33]	5	TB pleuritis, pericarditis, peritonitis	• Measurement of ADA levels in pleural, pericardial, and ascitic fluid has high sensitivity and specificity for extrapulmonary TB.
IFN-γ [13,15]	2	TB pleuritis	• Pleural fluid IFN-γ determination is a sensitive and specific test for the diagnosis of TB pleuritis.
Phage amplification assays [16]	1	Pulmonary TB	• Phage-based assays have high specificity but lower and variable sensitivity. • Their performance characteristics are similar to sputum microscopy.
Automated liquid cultures [10]	1	Pulmonary TB	• Automated liquid cultures are more sensitive than solid cultures. • Time to detection is more rapid than solid cultures.
Diagnosis of latent TB infection			
TST [34,35]	2	Latent TB infection	• Individuals who receive BCG vaccination are more likely to have a positive TST; the effect of BCG on TST results is less after 15 years; positive TST with indurations of >15 mm are more likely to be the result of TB infection than of BCG vaccination. • The effect on TST of BCG received in infancy is minimal, especially 10 years after vaccination. BCG received after infancy produces more frequent, more persistent, and larger TST reactions. • Non-tuberculous mycobacterial (NTM) infection is not a clinically important cause of false-positive TST, except in populations with a high prevalence of NTM sensitization and a very low prevalence of TB infection.
T cell–based IGRAs [21,26,37]	3	Latent TB infection	• IGRAs have excellent specificity (higher than the TST), and are unaffected by prior BCG vaccination.
Diagnosis of drug-resistant TB			
Phage amplification assays [25]	1	Rapid detection of rifampicin resistance	• When used on culture isolates, phage assays have high sensitivity, but variable and lower specificity. • In contrast, evidence is lacking on the accuracy of these assays when they are directly applied to sputum specimens.
Line probe assays: INNO-LiPA Rif. TB (LiPA) [22] and GenoType MTBDR assays [38]	2	Rapid detection of rifampicin resistance	• LiPA is a highly sensitive and specific test for the detection of rifampicin resistance in culture isolates, with relatively lower sensitivity when used directly on clinical specimens. •The GenoType MTBDR assays have excellent sensitivity and specificity for rifampicin resistance even when directly used on clinical specimens.
Colorimetric redox-indicator methods [19] and nitrate reductase assays [36]	2	Rapid detection of rifampicin and isoniazid resistance	• Colorimetric methods and nitrate reductase assays are highly sensitive and specific for the rapid detection of rifampicin and isoniazid resistance in culture isolates.

doi:10.1371/journal.pmed.0050156.t001

Diagnostic Accuracy) initiative was launched to improve the quality of reporting of diagnostic studies [60].

Conclusions

With the publication of several systematic reviews, there is now a strong evidence base to support global policy on TB diagnostics. A key challenge is to maintain the momentum gained in the past few years, and expand the scope and role of evidence synthesis to outcomes that go beyond conventional diagnostic accuracy. These outcomes include: accuracy of diagnostic algorithms (rather than single tests) and their relative contributions to the health care system; incremental or added value of new tests; impact of new tests on clinical decision-making and therapeutic choices; cost-effectiveness in routine programmatic settings; impact on patient-centered outcomes; and societal impact of new tools. Indeed, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to grading the quality of evidence and strength of recommendations for diagnostic tests recognizes that diagnostic accuracy results are surrogates for patient-centered outcomes, and emphasizes that diagnostic tests are of value only if they result in improved outcomes for patients [61].

In addition to expanding the scope of evidence synthesis, it is also important to ensure that systematic reviews stay current by including new literature. Periodic updates are needed to ensure that systematic reviews provide the most current evidence available for clinical and policy decisions. For example, the literature on IGRAs has exploded in the past few years, and this necessitated an updated meta-analysis on this topic [37].

Recognizing the growing importance of evidence-based TB diagnosis and policy making, the Stop TB Partnership's New Diagnostics Working Group has recently created a new subgroup on Evidence Synthesis for TB Diagnostics [62]. This subgroup will support the development of new systematic reviews, facilitate the development and dissemination of evidence summaries on new diagnostics, and actively promote their use in guideline and policy development processes, along the lines of the GRADE approach. ■

Acknowledgments

The authors acknowledge the excellent contributions made by authors of the systematic reviews cited in this work. Their efforts have made evidence-based TB diagnosis a reality. The authors are grateful to Professor S. P. Kalantri for helpful comments on a draft of this manuscript.

Funding: MP is a recipient of a New Investigator Career Award from the Canadian Institutes of Health Research. This agency had no involvement in the preparation of this manuscript.

Competing Interests: The authors have no financial conflicts of interest. All the authors are members of the Stop TB Partnership's Working Group on New Diagnostics. AR is the secretary of the Working Group. MP and ROB are co-chairs of the Working Group's subgroup on Evidence Synthesis for TB Diagnostics. ROB works for the Foundation for Innovative New Diagnostics, a nonprofit agency that collaborates with several industry partners for the development of new diagnostics for neglected infectious diseases. No industry partner was involved in the preparation of this manuscript.

References

1. Stop TB Partnership, World Health Organization (2006) The global plan to stop TB 2006–2015. Available: http://www.stoptb.org/globalplan/. Accessed 16 June 2008.
2. Raviglione MC, Uplekar MW (2006) WHO's new Stop TB Strategy. Lancet 367: 952-955.
3. World Health Organization (2007) New technologies for tuberculosis control: A framework for their adoption, introduction and implementation. Available: http://whqlibdoc.who.int/publications/2007/9789241595520_eng.pdf. Accessed 16 June 2008.
4. World Health Organization, Special Programme for Research and Training in Tropical Diseases (2006) Diagnostics for tuberculosis. Global demand and market potential. Available: http://www.who.int/tdr/publications/publications/tbdi.htm. Accessed 16 June 2008.
5. Oxman AD, Lavis JN, Fretheim A (2007) Use of evidence in WHO recommendations. Lancet 369: 1883-1889.
6. Hill S, Pang T (2007) Leading by example: A culture change at WHO. Lancet 369: 1842-1844.
7. World Health Organization (2008) Moving research findings into new WHO policies. Available: http://www.who.int/tb/dots/laboratory/policy/en/index4.html. Accessed 16 June 2008.
8. Pai M, McCulloch M, Enanoria W, Colford JM Jr. (2004) Systematic reviews of diagnostic test evaluations: What's behind the scenes? ACP J Club 141: A11-A13.
9. Daley P, Thomas S, Pai M (2007) Nucleic acid amplification tests for the diagnosis of tuberculous lymphadenitis: A systematic review. Int J Tuberc Lung Dis 11: 1166-1176.
10. Dinnes J, Deeks J, Kunst H, Gibson A, Cummins E, et al. (2007) A systematic review of rapid diagnostic tests for the detection of tuberculosis infection. Health Technol Assess 11: 1-196.
11. Flores LL, Pai M, Colford JM Jr., Riley LW (2005) In-house nucleic acid amplification tests for the detection of *Mycobacterium tuberculosis* in sputum specimens: meta-analysis and meta-regression. BMC Microbiol 5: 55.
12. Goto M, Noguchi Y, Koyama H, Hira K, Shimbo T, et al. (2003) Diagnostic value of adenosine deaminase in tuberculous pleural effusion: A meta-analysis. Ann Clin Biochem 40: 374-381.
13. Greco S, Girardi E, Masciangelo R, Capoccetta GB, Saltini C (2003) Adenosine deaminase and interferon gamma measurements for the diagnosis of tuberculous pleurisy: A meta-analysis. Int J Tuberc Lung Dis 7: 777-786.
14. Greco S, Girardi E, Navarra S, Saltini C (2006) The current evidence on diagnostic accuracy of commercial based nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis. Thorax 61: 783-790.
15. Jiang J, Shi HZ, Liang QL (2007) Diagnostic value of interferon-g in tuberculous pleurisy: A meta-analysis. Chest 131: 1133-1141.
16. Kalantri S, Pai M, Pascopella L, Riley L, Reingold A (2005) Bacteriophage-based tests for the detection of *Mycobacterium tuberculosis* in clinical specimens: A systematic review and meta- analysis. BMC Infect Dis 5: 59.
17. Liang QL, Shi HZ, Wang K, Qin SM, Qin XJ (2008) Diagnostic accuracy of adenosine deaminase in tuberculous pleurisy: A meta-analysis. Respir Med 102: 744-754.
18. Ling DI, Flores LL, Riley LW, Pai M (2008) Commercial nucleic-acid amplification tests for diagnosis of pulmonary tuberculosis in respiratory specimens: Meta-analysis and meta-regression. PLoS ONE 3: e1536. doi:10.1371/journal.pone.0001536
19. Martin A, Portaels F, Palomino JC (2007) Colorimetric redox-indicator methods for the rapid detection of multidrug resistance in *Mycobacterium tuberculosis:* A systematic review and meta-analysis. J Antimicrob Chemother 59: 175-183.
20. Mase SR, Ramsay A, Ng V, Henry M, Hopewell PC, et al. (2007) Yield of serial sputum specimen examinations in the diagnosis of pulmonary tuberculosis: A systematic review. Int J Tuberc Lung Dis 11: 485-495.
21. Menzies D, Pai M, Comstock G (2007) Meta-analysis: New tests for the diagnosis of latent tuberculosis infection: Areas of uncertainty and recommendations for research. Ann Intern Med 146: 340-354.
22. Morgan M, Kalantri S, Flores L, Pai M (2005) A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis:* A systematic review and meta-analysis. BMC Infect Dis 5: 62.
23. Pai M, Flores LL, Hubbard A, Riley LW, Colford JM Jr. (2004) Nucleic acid amplification tests in the diagnosis of tuberculous pleuritis: A systematic review and meta-analysis. BMC Infect Dis 4: 6.
24. Pai M, Flores LL, Pai N, Hubbard A, Riley LW, et al. (2003) Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: A systematic review and meta-analysis. Lancet Infect Dis 3: 633-643.
25. Pai M, Kalantri S, Pascopella L, Riley LW, Reingold AL (2005) Bacteriophage-based assays for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis:* A meta-analysis. J Infect 51: 175-187.
26. Pai M, Riley LW, Colford JM Jr. (2004) Interferon-gamma assays in the immunodiagnosis of tuberculosis: A systematic review. Lancet Infect Dis 4: 761-776.
27. Riquelme A, Calvo M, Salech F, Valderrama S, Pattillo A, et al. (2006) Value of adenosine deaminase (ADA) in ascitic fluid for the diagnosis of tuberculous peritonitis: A meta-analysis. J Clin Gastroenterol 40: 705-710.
28. Sarmiento OL, Weigle KA, Alexander J, Weber DJ, Miller WC (2003) Assessment by meta-analysis of PCR for diagnosis of smear-negative pulmonary tuberculosis. J Clin Microbiol 41: 3233-3240.
29. Steingart KR, Henry M, Laal S, Hopewell PC, Ramsay A, et al. (2007) A systematic review of commercial serological antibody detection

tests for the diagnosis of extra-pulmonary tuberculosis. Thorax 62: 911-918.
30. Steingart KR, Henry M, Laal S, Hopewell PC, Ramsay A, et al. (2007) Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: A systematic review. PLoS Med 4: e202. doi:10.1371/journal.pmed.0040202
31. Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, et al. (2006) Fluorescence versus conventional sputum smear microscopy for tuberculosis: A systematic review. Lancet Infect Dis 6: 570-581.
32. Steingart KR, Ng V, Henry M, Hopewell PC, Ramsay A, et al. (2006) Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: A systematic review. Lancet Infect Dis 6: 664-674.
33. Tuon FF, Litvoc MN, Lopes MI (2006) Adenosine deaminase and tuberculous pericarditis—A systematic review with meta-analysis. Acta Trop 99: 67-74.
34. Wang L, Turner MO, Elwood RK, Schulzer M, FitzGerald JM (2002) A meta-analysis of the effect of Bacille Calmette Guerin vaccination on tuberculin skin test measurements. Thorax 57: 804-809.
35. Farhat M, Greenaway C, Pai M, Menzies D (2006) False-positive tuberculin skin tests: What is the absolute effect of BCG and non-tuberculous mycobacteria? Int J Tuberc Lung Dis 10: 1192-1204.
36. Martin A, Panaiotov S, Portaels F, Hoffner S, Palomino JC, et al. (2008) The nitrate reductase assay for the rapid detection of isoniazid and rifampicin resistance in *Mycobacterium tuberculosis:* A systematic review and meta-analysis. J Antimicrob Chemother 62: 56-64.
37. Pai M, Zwerling A, Menzies D (2008) Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection—An update. Ann Intern Med 149: 177-184.
38. Ling DI, Zwerling A, Pai M (2008) GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: A meta-analysis. Eur Respir J. E-pub 9 July 2008. doi:10.1183/09031936.00061808
39. Andersen P, Doherty TM, Pai M, Weldingh K (2007) The prognosis of latent tuberculosis: Can disease be predicted? Trends Mol Med 13: 175-182.
40. Trajman A, Pai M, Dheda K, van zyl Smit R, Zwerling A, et al. (2008) Novel tests for diagnosing tuberculous pleural effusion: What works and what does not? Eur Respir J 31: 1098-1106.
41. World Health Organization (2007) The global MDR-TB and XDR-TB response plan 2007–2008. Available: http://whqlibdoc.who.int/hq/2007/WHO_HTM_TB_2007.387_eng.pdf. Accessed 16 June 2008.
42. Hopewell PC, Pai M, Maher D, Uplekar M, Raviglione MC (2006) International standards for tuberculosis care. Lancet Infect Dis 6: 710-725.
43. Getahun H, Harrington M, O'Brien R, Nunn P (2007) Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings: Informing urgent policy changes. Lancet 369: 2042-2049.
44. World Health Organization (2007) Reduction of number of smears for the diagnosis of pulmonary TB. Available: http://www.who.int/tb/dots/laboratory/policy/en/index2.html. Accessed 16 June 2008.
45. World Health Organization (2007) Definition of a new sputum smear-positive TB case. Available: http://www.who.int/tb/dots/laboratory/policy/en/index1.html. Accessed 16 June 2008.
46. Squire SB, Belaye AK, Kashoti A, Salaniponi FM, Mundy CJ, et al. (2005) 'Lost' smear-positive pulmonary tuberculosis cases: Where are they and why did we lose them? Int J Tuberc Lung Dis 9: 25-31.
47. World Health Organization (2007) The use of liquid medium for culture and DST. Available: http://www.who.int/tb/dots/laboratory/policy/en/index3.html. Accessed 16 June 2008.
48. Ridderhof JC, van Deun A, Kam KM, Narayanan PR, Aziz MA (2007) Roles of laboratories and laboratory systems in effective tuberculosis programmes. Bull World Health Organ 85: 354-359.
49. World Health Organization, Special Programme for Research and Training in Tropical Diseases (2006) Request for applications: Diagnostic trial sites: Improving the diagnosis of tuberculosis through optimization of sputum smear microscopy. Available: http://www.who.int/tdr/grants/grants/trial_sites.htm. Accessed 16 June 2008.
50. Foundation for Innovative New Diagnostics (2008) Market launch of improved fluorescence microscope scheduled for later this year. Available: http://www.finddiagnostics.org/news/press/zeiss_mar08.pdf. Accessed 16 June 2008.
51. Cunningham J (2005) Rapid serological-based TB test evaluation: Prelim analysis. Stop TB Working Group on New Diagnostics. Available: http://www.stoptb.org/wg/new_diagnostics/assets/documents/jane_cunningham.pdf. Accessed 16 June 2008.
52. Perkins MD, Cunningham J (2007) Facing the crisis: Improving the diagnosis of tuberculosis in the HIV era. J Infect Dis 196 (Suppl 1): S15-S27.
53. Canadian Tuberculosis Committee (2007) Interferon gamma release assays for latent tuberculosis infection. An Advisory Committee Statement (ACS). Can Commun Dis Rep 33: 1-18.
54. HPA Tuberculosis Programme Board (2007) Health Protection Agency Position Statement on the use of Interferon Gamma Release Assay (IGRA) tests for tuberculosis (TB): Draft for consultation. Available: http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1204186168242. Accessed 16 June 2008.
55. National Tuberculosis Advisory Committee Australia (2007) Position statement on interferon-gamma release immunoassays in the detection of latent tuberculosis infection, October 2007. Commun Dis Intell 31: 404-405.
56. Pai M, Dheda K, Cunningham J, Scano F, O'Brien R (2007) T-cell assays for the diagnosis of latent tuberculosis infection: Moving the research agenda forward. Lancet Infect Dis 7: 428-438.
57. Pai M, O'Brien R (2006) Tuberculosis diagnostics trials: Do they lack methodological rigor? Expert Rev Mol Diagn 6: 509-514.
58. Small PM, Perkins MD (2000) More rigour needed in trials of new diagnostic agents for tuberculosis. Lancet 356: 1048-1049.
59. Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, et al. (2006) Evaluation of diagnostic tests for infectious diseases: General principles. Nat Rev Microbiol 4: S21-S31.
60. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 138: 40-44.
61. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, et al. (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 336: 1106-1110.
62. Stop TB Partnership Working Group on New TB Diagnostics (2007) Revised strategic plan of the Stop TB Partnership's Working Group on New Diagnostics. Available: http://www.stoptb.org/wg/new_diagnostics/assets/documents/Draft%20NDWG%20Strategic%20Plan%20for%20Cape%20Town%20Meeting.pdf. Accessed 16 June 2008.

New and improved tuberculosis diagnostics: evidence, policy, practice, and impact

Madhukar Pai[a], Jessica Minion[a], Karen Steingart[b] and Andrew Ramsay[c]

[a]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, [b]Francis J. Curry National Tuberculosis Center, University of California, San Francisco, California, USA and [c]UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR), World Health Organization, Geneva, Switzerland

Correspondence to Madhukar Pai, MD, PhD, Assistant Professor, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, QC H3A 1A2, Canada
Tel: +1 514 398 5422; fax: +1 514 398 4503;
e-mail: madhukar.pai@mcgill.ca

Current Opinion in Pulmonary Medicine 2010, 16:271–284

Purpose of review

The aim is to summarize the evidence base for tuberculosis (TB) diagnostics, review recent policies on TB diagnostics, and discuss issues such as how evidence is translated into policy, limitations of the existing evidence base, and challenges involved in translating policies into impact.

Recent findings

Case detection continues to be a major obstacle to global TB control. Fortunately, due to an unprecedented level of interest, funding, and activity, the new diagnostics pipeline for TB has rapidly expanded. There have been several new policies and guidelines on TB diagnostics. However, there are major gaps in the existing pipeline (e.g. lack of a point-of-care test) and the evidence base is predominantly made up of research studies of test accuracy.

Summary

With the availability of new diagnostics and supporting policies, the next major step is translation of policy into practice. The impact of new tests will depend largely on the extent of their introduction and acceptance into the global public sector. This will itself depend in part on policy decisions by international technical agencies and national TB programs. With the engagement of all key stakeholders, we will need to translate evidence-based policies into epidemiological and public health impact.

Keywords

diagnostics, evidence, impact, policy, tuberculosis

Curr Opin Pulm Med 16:271–284

1070-5287

Introduction

In 2010, poor diagnosis remains a major obstacle to global tuberculosis (TB) control. In most high-burden countries, TB is still diagnosed using tools such as direct sputum microscopy and chest radiographs. Fortunately, the past few years have seen an unprecedented level of interest, funding support, and activity focused on the development of new tools for TB diagnosis, and the new diagnostics pipeline for TB is rapidly expanding. In parallel, there have been several new policy recommendations on TB diagnostics by the WHO. Because recent publications [1•,2,3•,4] have exhaustively reviewed the current pipeline of new diagnostics and the expanding evidence base for their use, we focus our attention on how evidence is translated into policy, limitations of the existing evidence base, deficiencies in the current diagnostics pipeline, and challenges involved in translating policies into practice and impact.

What is the evidence base for tuberculosis diagnostics?

The evidence base for TB diagnostics is ultimately derived from a large body of original research. Because individual studies are seldom sufficient to inform policy and guideline development, the totality of available evidence must be synthesized. Thus, systematic reviews and meta-analyses are often necessary to summarize the evidence on a given diagnostic test. In the past decade, there have been over 35 systematic reviews published on TB diagnostics, on topics ranging from smear microscopy to molecular diagnostics and in-vitro assays for latent TB infection (LTBI). All of these systematic reviews have been made available on a new website 'Evidence-based Tuberculosis Diagnosis' (www.tbevidence.org) compiled by the Stop TB Partnership's New Diagnostics Working Group, in collaboration with several agencies [5•]. While the key findings of published systematic reviews and meta-analyses on TB diagnostics have been reviewed elsewhere [6•], Table 1 provides a brief overview of the evidence base for TB diagnosis, essentially synthesizing the evidence from several systematic reviews [7–37].

What is lacking in current evidence base?

Although a large number of systematic reviews have been published on TB diagnostics, almost all focus on test accuracy (i.e. sensitivity and specificity). This is in part

Table 1 Summary of findings from systematic reviews on tuberculosis diagnostic tests

Diagnostic test	Description	Disease/site	Major findings/results of systematic reviews	Major references
Diagnosis of active TB				
Sputum smear microscopy	Microscopic observation of stained acid fast bacilli	Pulmonary TB	Fluorescence microscopy (FM) is on average 10% more sensitive than conventional microscopy. Specificity of both fluorescence and conventional microscopy is similar. Fluorescent microscopy is associated with improved time efficiency. LED FM performs equivalently to conventional FM, with added benefits of low cost, durability, and ability to use without a darkroom. Centrifugation and overnight sedimentation, preceded with any of several chemical methods (including bleach) is slightly more sensitive (6–9%) than direct microscopy; specificity may be slightly decreased (1–3%) by sputum processing methods. When serial sputum specimens are examined, the mean incremental yield and/or increase in sensitivity from examination of third sputum specimen ranges between 2 and 5%.	[7–9]
Nucleic acid amplification tests (NAATs)	Isolation, replication, and detection of nucleic acid sequences specific for *Mycobacterium tuberculosis*	Pulmonary and extrapulmonary TB	NAATs have high specificity and positive predictive value. NAATs, however, have relatively lower (and highly variable) sensitivity and negative predictive value for all forms of TB, especially in smear-negative and extrapulmonary disease. In-house ('home brew') NAATs produce highly inconsistent results as compared with commercial, standardized NAATs.	[10–15]
Commercial serological antibody detection tests	Detection of host antibody response to *Mycobacterium tuberculosis* antigens	Pulmonary and extrapulmonary TB	Serological tests for both pulmonary and extrapulmonary TB produce highly inconsistent estimates of sensitivity and specificity; none of the current assays perform well enough to replace microscopy.	[12,16,17]
Noncommercial (in-house) serological antibody detection tests	Detection of host antibody response to *Mycobacterium tuberculosis* antigens	Pulmonary TB	Several potential candidate antigens for inclusion in an antibody detection-based diagnostic test for pulmonary TB in HIV-infected and -uninfected individuals were identified. Combinations of select antigens provide higher sensitivities than single antigens.	[18]
Adenosine deaminase (ADA)	Detection of host cellular enzyme released by lymphocytes in response to live intracellular pathogens	TB pleuritis, pericarditis, peritonitis	Measurement of ADA levels in pleural, pericardial, and ascitic fluid is a useful adjunct test for TB pleuritis, pericarditis, and peritonitis. Systematic reviews have reported pooled sensitivity estimates between 88 and 100%, and specificity estimates between 83 and 97%.	[19,20]
Interferon-gamma (IFN-γ)	Measurement of IFN-γ	TB pleuritis	Pleural fluid IFN-γ determination appears to be a useful diagnostic for TB pleuritis, with systematic reviews reporting pooled sensitivity estimates between 89 and 96%, and specificity estimates between 96 and 97%.	[19,21]
Phage amplification assays	Detection of *Mycobacterium tuberculosis*-specific phage viruses, after their infection and amplification of live MTB	Pulmonary TB	Despite high-accuracy estimates, current phage-based assays are limited by high rates of indeterminate results (up to 36%).	[22]
Automated liquid cultures	Automated detection of changes in oxygen, carbon dioxide, or pressure resulting from bacterial growth	Pulmonary TB	Automated liquid cultures are more sensitive than solid cultures; time to detection is more rapid than solid cultures.	[12,23]
Diagnosis of latent TB				
Tuberculin skin test (TST)	Measurement of induration as a result of exposure to intradermal tuberculin	Latent TB infection	Individuals who have received BCG vaccination are more likely to have a positive TST; the effect of BCG on TST results is less after 15 years; positive TST with indurations of >15 mm are more likely to be the result of TB infection than of BCG vaccination. The effect on TST of BCG received in infancy is minimal, especially 10 years after vaccination. BCG received after infancy produces more frequent, more persistent, and larger TST reactions. Nontuberculous mycobacterial (NTM) infection is not a clinically important cause of false-positive TST, except in populations with a high prevalence of NTM sensitization and a very low prevalence of TB infection.	[24,25]
T-cell-based interferon-γ release assays (IGRAs)	Measurement of IFN-γ released from lymphocytes when stimulated by *Mycobacterium tuberculosis*-specific antigens	Latent TB infection	IGRAs have excellent specificity (higher than the tuberculin skin test) and are unaffected by prior BCG vaccination. IGRAs cannot distinguish between LTBI and active TB and have no role for active TB diagnosis in adults. Used as an adjunctive diagnostic, IGRAs may aid in the investigation of pediatric TB. IGRAs correlate well with markers of TB exposure in low-incidence countries. IGRA performance appears to differ in high-endemic vs. low-endemic countries. IGRA sensitivity varies across populations and tends to be lower in high-endemic countries and in HIV-infected individuals.	[12,26–29]

Diagnosis of drug resistance				
Phage amplification assays	Detection of *Mycobacterium tuberculosis*-specific phage viruses, after their infection and amplification of live MTB + inhibition of growth in presence of antituberculous drugs	Rapid detection of rifampicin resistance	When used on culture isolates, phage assays have high sensitivity, but variable and lower specificity. In contrast, evidence is lacking on the accuracy of these assays when they are directly applied to sputum specimens. Recent studies have raised concerns about contamination, false-positive results, and technical assay failures.	[30,31]
Line probe assays: INNO-LiPA Rif. TB [LiPA] and GenoType MTBDR assay	Detection of genetic sequences associated with resistance (after extraction and amplification) using immobilized probes and colorimetric development	Rapid detection of rifampicin resistance	LiPA is a highly sensitive and specific test for the detection of rifampicin resistance in culture isolates. The test has relatively lower sensitivity when used directly on clinical specimens. The GenoType MTBDR assays have excellent sensitivity and specificity for rifampicin resistance even when directly used on clinical specimens.	[32–34]
Colorimetric redox indicators (CRIs)	Determination of MIC using microdilution, followed by addition of reagent which will become reduced in the presence of actively growing MTB resulting in a color change	Rapid detection of rifampicin and isoniazid resistance	Colorimetric methods are sensitive and specific for the detection of rifampicin and isoniazid resistance in culture isolates. CRIs use inexpensive noncommercial supplies and equipment and have a rapid turnaround time (7 days).	[35]
Nitrate reductase assays (NRAs)	Direct or indirect inoculation of drug-free and drug-containing media containing KNO_3. Addition of Greiss reagent detects early growth by reacting with enzymatic byproduct and resulting in a color change.	Rapid detection of rifampicin and isoniazid resistance	NRA has high accuracy when used to detect rifampicin and isoniazid resistance in culture isolates. Limited data are available on its use when directly applied to clinical specimens, but results are promising. The NRA is simple, uses inexpensive noncommercial supplies and equipment, and has a rapid turnaround time (7–14 days) compared to conventional methods.	[36]
Microscopic observation drug susceptibility (MODS)	Direct or indirect inoculation of drug-free and drug-containing liquid media, followed by examination using an inverted microscope to detect early growth	Rapid detection of rifampicin and isoniazid resistance	MODS has high accuracy when testing for rifampicin resistance, but shows slightly lower sensitivity when detecting isoniazid resistance. MODS appears to perform equally well using direct patient specimens and culture isolates. MODS uses noncommercial supplies and equipment, and has a rapid turnaround time (10 days) compared with conventional methods.	[37]
Thin layer agar (TLA)	Direct or indirect inoculation of drug-free and drug-containing solid media, followed by examination using a microscope to detect early growth	Rapid detection of rifampicin and isoniazid resistance	There is a paucity of data evaluating TLA for the detection of drug susceptibility; however, all studies to date have found 100% concordance with their reference standards. TLA uses inexpensive noncommercial supplies and equipment, and has a rapid turnaround time (11 days) compared with conventional methods.	[37]

BCG, bacillus Calmette-Guérin; LED, light emitting diode; LTBI, latent TB infection; MIC, minimal inhibitory concentration; MTB, *Mycobacterium* tuberculosis; TB, tuberculosis. Adapted from [6•]. (Open Access under Creative Commons Attribution License).

because a large proportion of TB diagnostic research studies are focused on measuring test accuracy. Findings from systematic reviews suggest that even relatively straightforward studies of test accuracy are often poorly designed and reported [38,39]. Both researchers of primary TB diagnostic studies and authors of systematic reviews and meta-analyses need to make efforts to follow published guidelines for conducting and reporting their work [40,41], to make the most of their contribution to a useful and unbiased literature base.

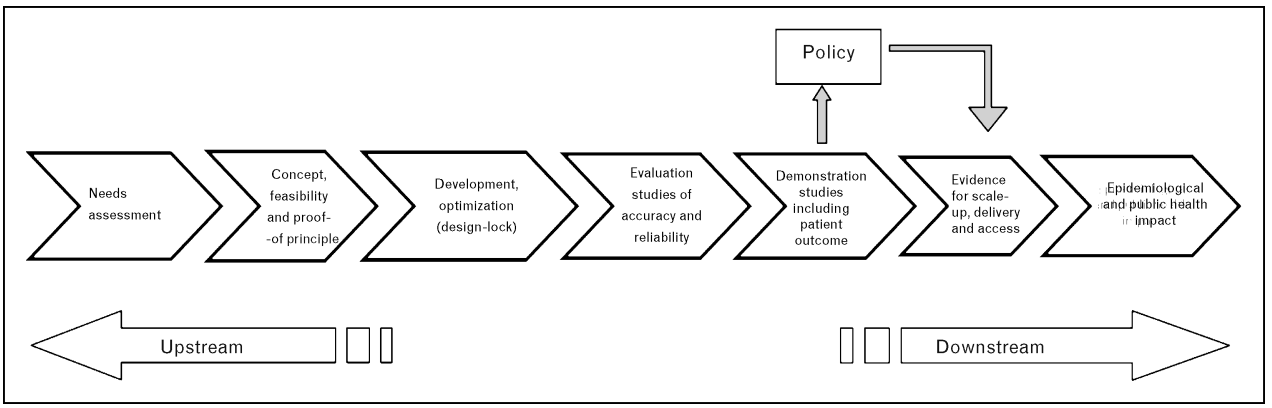
Although the quality of diagnostic studies measuring test accuracy is important, evidence about test accuracy is only one link in a long chain of activities that make up the pathway to developing and implementing a new TB diagnostic. In 2009, the Stop TB Partnership's New Diagnostics Working Group published a scientific blueprint for development of new TB diagnostics [42••]. This publication provides a comprehensive, well referenced plan to guide researchers, clinicians, industry partners, academics, and TB controllers in all sectors in all aspects of TB diagnostics development [42••], starting from needs' assessment, concept, feasibility, proof-of-concept, to test development, validation, and, ultimately, delivery, scale-up, access, and epidemiological and public health impact.

As shown in Fig. 1, evidence on test accuracy is essential, but policy development requires more than estimation of test accuracy. Along with data on test accuracy, we need to consider user-important as well as patient-important outcomes. Patient-important outcomes require more sophisticated and often more resource-intensive research [43,44], wherein a study shows that implementing a diagnostic test in a given situation results in clinically relevant improvements in patient care and/or patient outcomes. For TB diagnostics, this might mean an increased number of patients detected and receiving appropriate treatment, fewer patients defaulting from the diagnostic pathway due to reduced number of patient visits, or more patients cured due to accurate detection of drug resistance. Studies may also investigate the values and preferences patients have when choosing one diagnostic test compared to another. Although the challenges and costs of demonstrating these types of outcomes make them unattractive for many researchers and funding agencies, it is no less important than proving a therapeutic intervention actually changes the course of a disease and not just the level of a biomarker or surrogate endpoint.

User-important outcomes consist of practical concerns for the usability of a test in real-world situations. Although these generally do not require fundamentally different strategies to evaluate, it is important that they are assessed under implementation or real-world settings. These include the ease of use of a technology, the hands-on time of performing the test, the expertise or training required, and the infrastructure needed. It is important to consider biosafety, robustness of any equipment involved, as well as pragmatic issues such as the shelf-life of reagents, the need for special shipping or storage of materials, the availability and reliability of supply chains, and of course cost.

These types of evidence must be taken into account, along with test accuracy and reliability, when policy makers or programs are evaluating a diagnostic for recommendation or widespread use. Systematic reviews of diagnostics should make an effort to summarize data on these outcomes in addition to accuracy, appraise the quality of available evidence, and explore the uncertainty regarding the often assumed values and preferences of patients associated with these tests. However,

Figure 1 Level of evidence required for policy process



Adapted from [42••].

an obstacle here is a lack of methodology for collecting and analyzing such evidence even if the data were reported in primary research. In other words, currently used systematic review methods are mainly aimed at test accuracy.

Where is the current diagnostics pipeline deficient?

Although there are many more TB diagnostics in the pipeline today than in the past, the existing TB diagnostics pipeline itself has limitations and neglects some important aspects of the TB epidemic. Table 2 summarizes the major research priorities for TB diagnostics.

The biggest concern continues to be the lack of a rapid, simple, inexpensive, point-of-care (POC) test for active TB. As yet nothing has emerged from the pipeline or looks likely to emerge from the pipeline in the near future that could supplant smear microscopy. An easy-to-use, inexpensive diagnostic that can perform as well or better than smear microscopy and can deliver results within minutes without sophisticated equipment or highly-trained laboratory personnel would be a major step forward in TB diagnostics and could have a tremendous impact on global TB control [45,46•].

Another area still lacking in adequate diagnostic options is smear-negative TB, especially in HIV-infected persons [47]. Undiagnosed TB is very common in persons infected with HIV; therefore, intensive active case finding is required as strategies that rely on passive detection, or screening with smear microscopy alone, will miss a large number of coinfected patients [47]. Considering the proven benefit of TB preventive therapy using isoniazid in HIV-infected persons, ruling out active TB before initiation of single drug treatment is important not only for the care of the individual patient, but also to prevent the inadvertent selection for drug resistance. The development and validation of an algorithm, taking advantage of newly available tests, to aggressively target this high-risk population remains a priority for TB control.

Childhood TB presents similar challenges [48]. By virtue of the pathophysiology of TB in pediatrics and the inability to obtain adequate sputum samples, microbiologic confirmation of active TB remains an insensitive and inadequate standard. Similar to patients with HIV and smear-negative TB, the development and improvement in diagnostic algorithms that take advantage of available new diagnostics is needed. As good quality sputum specimens are difficult to collect, novel diagnostics that can be used on urine, saliva, breath condensate, and so on could have a greater impact in these populations, especially if a POC format could be developed.

The control of drug-resistant TB requires accurate and rapid diagnostics for the detection of critical patterns of drug resistance. The need to identify cases of multidrug-resistant TB (MDR-TB) through detecting resistance to rifampicin and isoniazid is now well recognized. The next step is to accurately and rapidly identify cases of extensively drug-resistant TB (XDR-TB) through the detection of resistance to key second-line drugs.

Although new tests [such as interferon-gamma release assays (IGRAs)] have emerged for LTBI diagnosis, these tests cannot resolve the various phases of the latent TB spectrum [49,50]. This means existing tests cannot be used to target preventive therapy at the subgroup that is most likely to benefit from treatment. Thus, there is a need for a highly predictive biomarker or combination of biomarkers, which will allow accurate prediction of the subgroup of latently infected individuals who are at highest risk of progression to disease.

How is evidence translated into policy?

The WHO has taken the lead on developing policies and guidelines on TB diagnostics. The WHO policy process is summarized in a recent statement entitled 'Moving research findings into new WHO policies [51•].' The key steps in the WHO policy process are given in Table 3 [51•]. This process takes into account the importance of not only identifying areas in need of policy guidance, but also ensuring that policies are evidence-based and then followed up with dissemination and promotion of new recommendations. For step 2, reviewing the evidence, WHO may commission a systematic review and meta-analysis of available data (published and unpublished) using standard methods appropriate for diagnostic accuracy studies [52•].

Table 4 provides an overview of all the recent WHO policies on TB diagnostics [51•,53–57]. Since 2007, the WHO has endorsed several diagnostic tests and strategies, including liquid cultures, optimized smear microscopy, line probe assays, and noncommercial culture systems for drug-susceptibility testing.

The foundation of the WHO policy process is now the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [58••]. This is in part a response to the criticism that systematic reviews are rarely used for developing WHO recommendations and that WHO policy processes usually rely heavily on expert opinion [59]. The GRADE approach provides a system for rating the quality of evidence and the strength of recommendations that is explicit, comprehensive, transparent, and pragmatic and is being adopted increasingly by organizations worldwide [58••,60]. The WHO now requires the use of GRADE for

Table 2 Major priorities for research and development and implementation of tuberculosis diagnostics

Research priority	Research methods	Expected outcome	Justification
Development of a rapid, accurate point-of-care (POC) test for pulmonary TB.	Biomarker discovery, followed by incorporation in a highly sensitive POC platform and then clinical validation.	A POC test for pulmonary TB that will meet the user-defined specifications (such as those proposed by MSF).	Currently, there is no POC test for TB that can be used at the health clinic level. Diagnostic delays, therefore, are common.
Development and validation of tools for rapid detection of drug resistance, including for XDR-TB and standardization of DST for second-line drugs.	Identification and characterization of mutations associated with second-line drug resistance; development of newer generation molecular assays for MDR/XDR-TB; improved standardization of existing tests for second-line DST.	Rapid molecular (genotypic) assays for MDR/XDR-TB that will allow rapid identification of drug-resistant TB.	Although line-probe assays are highly accurate for rifampicin resistance, accuracy is lower for isoniazid and other drugs. Second-line DST continues to be a challenge; mutations are not well defined and standardization is a problem with phenotypic methods.
Intensified, active case detection strategies for early detection of active TB in HIV-infected persons (at the clinic level and in the community).	Development and validation of an algorithm (including new tests) for rapid detection of TB in HIV-infected persons.	A validated algorithm that will enable detection of TB in a large proportion of HIV-infected persons with TB disease.	Passive case detection methods do not work well in areas with high HIV prevalence; undiagnosed TB is frequent in HIV-infected persons and can cause enormous morbidity and mortality. Aggressive case detection approaches are needed to enhance case detection, reduce mortality, and reduce transmission.
Improving current diagnostic algorithms to shorten the time required for establishing a diagnosis of smear-negative pulmonary TB and extrapulmonary TB in HIV-infected persons and children.	Development and validation of an algorithm (including new tests) for rapid detection of smear-negative and extrapulmonary TB in HIV-infected persons and children.	A validated algorithm that will rapidly enable detection of smear-negative and extrapulmonary TB in a large proportion of HIV-infected persons and children.	Smear-negative TB, extrapulmonary TB, and childhood TB are diagnostic challenges and available tests perform poorly in these cases of paucibacillary TB. Newer algorithms and tests are needed to get around the limitations of current methods.
Development of a test or algorithm that can accurately rule out active TB disease in HIV-infected persons [to allow initiation of preventive therapy (IPT)]	Development and validation of an algorithm (including new tests) for rapidly ruling out active TB (including smear-negative and extrapulmonary TB) in HIV-infected persons.	A validated algorithm that will enable exclusion of TB in a large proportion of HIV-infected persons prior to IPT.	In HIV-infected persons, undiagnosed active TB is common. Before IPT, it is necessary to rule out active TB. However, there is no easy and accurate method to do this in high-burden countries.
Which biomarkers or combinations of markers will help distinguish the various stages of the spectrum of latent TB infection (from sterilizing immunity to subclinical active disease) and will allow accurate prediction of the subgroup of latently infected individuals who are at highest risk of progression to disease.	Biomarker discovery, followed by validation in clinical and longitudinal (cohort) studies for markers that can predict risk of progression to active TB.	Identification of a biomarker or combination of biomarkers that will allow accurate prediction of the subgroup of latently infected individuals who are at highest risk of progression to disease.	Existing tests for latent TB (TST and IGRAs) cannot resolve the various phases of the latent TB spectrum. This means existing tests cannot be used to target IPT at the subgroup that is most likely to benefit from treatment. This results in overtreatment of a large number of latently infected persons.
Development of a rapid test for childhood TB that will not depend on sputum specimen testing.	Development and validation of a test or an algorithm (including new tests) for rapid detection of TB in children, without requiring sputum specimens.	A test (preferably POC) that can use nonsputum specimens (e.g. urine or breath condensate or saliva) for rapid detection of TB in children.	Childhood TB is a diagnostic challenge and available tests perform poorly in these cases of paucibacillary TB. Also, as young children are unable to produce sputum, it will be helpful to use alternative specimens such as urine, saliva, or breath condensate.
Define different ways of operationalizing and implementing existing policies on HIV testing of TB patients and TB screening of HIV-infected persons.	Operational research on different ways of implementing existing policies on HIV testing of TB patients and TB screening of HIV-infected persons.	Identification of at least one feasible approach that might work best and therefore can be scaled-up.	Existing policies on HIV testing of TB patients and TB screening of HIV-infected persons are poorly implemented. A large proportion of TB patients are not tested for HIV, and HIV-infected persons are not screened for TB. This results in undiagnosed co-infection morbidity/mortality and continued transmission in the community.
Once new diagnostics are approved and available, what factors can enhance their actual delivery and implementation at the programmatic level in high-burden countries?	Operational research on different ways of implementing new diagnostics in national TB programs in high-burden settings.	Identification of at least one feasible implementation approach that might work best and therefore can be scaled-up.	Availability of new tools does not necessarily ensure their adoption and implementation. Translation of policy into practice requires better understanding of barriers to implementation and methods to overcome such barriers.
What is the likely epidemiological impact of widespread LTBI diagnosis and treatment in high-burden countries, and what contribution will LTBI diagnosis and treatment make toward the attainment of the Stop TB Partnership's target for TB elimination?	Mathematical modeling study.	The modeling study will inform the debate on when high-burden countries should begin to focus attention on LTBI diagnosis and treatment.	LTBI diagnosis and treatment is currently not a priority in high-burden countries. However, as TB incidence falls, it can become a priority. Also, some recent modeling studies suggest that TB elimination will require strategies aimed at LTBI management.

DST, drug susceptibility testing; LTBI, latent tuberculosis infection; MDR, multidrug-resistant; MSF, Médecins Sans Frontières; XDR, extensively drug-resistant.

Table 3 World Health Organization policy process for tuberculosis

Major steps	Description of the process
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 and 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.
Reviewing the evidence (including systematic reviews)	WHO may carry out or commission a review of the documentation of 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.
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 (using the GRADE approach) and make a recommendation or propose draft policies or guidelines to WHO's Strategic and Technical Advisory Group for Tuberculosis (STAG-TB).
Assessing draft policies and guidelines	STAG-TB provides objective, ongoing technical, and strategic advice to WHO related to TB care and control. STAG-TB's objectives are to provide the Director-General, through the Stop TB Department, 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.
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 website, list serves, and journal publications. WHO also disseminates its recommendations to other agencies and donors engaged in TB control activities.

GRADE, Grading of Recommendations Assessment, Development and Evaluation; TB, tuberculosis. World Health Organization: moving research findings into new WHO policies [51•].

all new and revised WHO policies and guidelines, including policies on diagnostics [61]. For example, recent WHO policies on TB infection control [62] and the revised TB treatment guidelines [63] used the GRADE approach.

Grading of Recommendations Assessment, Development, and Evaluation for diagnostic tests: strengths and limitations

The GRADE approach provides a clear separation of quality of evidence and strength of recommendations [58••]. In judgments about quality of evidence, GRADE considers six factors: study design, methodological quality, directness of evidence (patient-important outcomes and generalizability), inconsistency of results, imprecision of results (imprecise or sparse data), and publication bias [58••]. Thus, quality of evidence reflects our confidence that estimates of benefits and downsides from a diagnostic test or strategy generated from research are correct. Quality of evidence is graded as follows:

(1) High quality: further research is very unlikely to change our confidence in the estimate of effect.
(2) Moderate quality: further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
(3) Low quality: further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
(4) Very low quality: any estimate of effect is very uncertain.

In the GRADE approach, well designed studies of diagnostic accuracy (cross-sectional or cohort studies on patients with diagnostic uncertainty and use of appropriate reference standard) can provide high-quality evidence on test accuracy. However, these studies may provide only low-quality evidence for guideline development because of uncertainty about the link between test accuracy and outcomes important to patients (discussed below).

The strength of a recommendation refers to the extent to which one can be confident that adherence to the recommendation will do more good than harm [58••]. There are four factors to consider: balance between desirable and undesirable effects; quality of evidence; values and preferences; and costs (resource allocation). GRADE classifies recommendations as strong (most informed patients would choose this option) or weak (patients' choices will vary according to their values and preferences and not all patients would choose this option).

The GRADE process was initially developed for treatment interventions and, therefore, tends to be focused on

Table 4 Highlights of recent WHO policies and statements on tuberculosis diagnostics

Year policy was made	Purpose of testing	Diagnostic test or approach	WHO recommendations
2007	Case detection and drug-susceptibility testing (DST)	Liquid media for culture and DST	WHO recommends, as a step-wise approach: The use of liquid medium for culture and DST in middle-income and low-income countries. The rapid species identification to address the needs for culture and DST. Taking into consideration that liquid systems will be implemented in a phased manner, integrated into a country-specific comprehensive plan for laboratory capacity strengthening.
2007	Case detection	Definition of a new sputum smear-positive TB case	The revised definition of a new sputum smear-positive pulmonary TB case is based on the presence of at least one acid fast bacilli (AFB+) in at least one sputum sample in countries with a well functioning external quality assurance (EQA) system.
2007	Case detection	Reduction of number of smears for the diagnosis of pulmonary TB	WHO recommends the number of specimens to be examined for screening of TB cases can be reduced from three to two, in places where a well functioning EQA system exists, where the workload is very high and human resources are limited.
2008	DST	Molecular line probe assays for rapid screening of patients at risk of MDR-TB	The use of line probe assays is recommended by WHO, with the following guiding principles: Adoption of line probe assays for rapid detection of MDR-TB should be decided by Ministries of Health within the context of country plans for appropriate management of MDR-TB patients, including the development of country-specific screening algorithms and timely access to quality-assured second-line anti-TB drugs. Line probe assay performance characteristics have been adequately validated in direct testing of sputum smear-positive specimens and on isolates of *Mycobacterium tuberculosis* complex grown from smear-negative and smear-positive specimens. Direct use of line probe assays on smear-negative clinical specimens is not recommended. The use of commercial line probe assays, rather than in-house assays, is recommended to ensure reliability and reproducibility of results, as in-house assays have not been adequately validated or used outside limited research settings. Adoption of line probe assays does not eliminate the need for conventional culture and DST capability; culture remains necessary for definitive diagnosis of TB in smear-negative patients, whereas conventional DST is required to diagnose extensively drug-resistant TB (XDR-TB). As current line probe assays only detect resistance to rifampicin and/or isoniazid, countries with documented or suspected cases of XDR-TB should establish or expand conventional culture and DST capacity for quality-assured susceptibility testing of second-line drugs, based on current WHO policy guidance.
2009	Case detection	LED-based microscopy	WHO recommends that conventional fluorescence microscopy be replaced by LED microscopy in all settings and that LED microscopy be phased in as an alternative for conventional ZN microscopy in both high-volume and low-volume laboratories. The switch to LED microscopy should be carried out through a carefully phased implementation plan, using LED technologies that meet WHO specifications.
2009	DST	Noncommercial culture and DST methods	WHO recommends that selected noncommercial culture and DST methods be used as an interim solution in resource-constrained settings, in reference laboratories, or those with sufficient culture capacity, while capacity for genotypic and/or automated liquid culture and DST are being developed. With due consideration of the above issues, WHO endorses the selective use of one or more of the following noncommercial culture and DST methods: Microscopically observed drug susceptibility (MODS), for rapid screening of patients suspected of having MDR-TB, under clearly defined programmatic and operation conditions, and once speciation concerns have been adequately addressed without compromising bio-safety; The nitrate reductase assay (NRA), for screening of patients suspected of having MDR-TB, under clearly defined programmatic and operation conditions, and acknowledging that time to detection of MDR in indirect application would not be faster (but less expensive) than conventional DST methods using commercial liquid culture or line probe assays; Colorimetric redox indicator (CRI) methods, as indirect tests on *M. tuberculosis* isolates from patients suspected of having MDR-TB, under clearly defined programmatic and operation conditions, and acknowledging that time to detection of MDR would not be faster (but less expensive) than conventional DST methods using commercial liquid culture or line probe assays.

LED, light-emitting diode; MDR-TB, multidrug-resistant tuberculosis. From the World Health Organization [51•,53–56].

randomized controlled trials (RCTs). It has been adapted for diagnostic tests and strategies [64••,65], though this area is a work in progress and can be improved based on user's feedback. The first time the GRADE approach was applied to TB diagnostics by the WHO was in September 2009 for use in developing guidelines for improving sputum smear microscopy and using noncommercial culture methods for rapid detection of TB drug resistance. From these experiences, we have found the GRADE approach to have several strengths as well as some limitations.

On the positive side, GRADE offers a systematic, objective, and transparent process and requires the explicit use of systematic reviews and evidence summaries. GRADE forces us to consider several elements, including quality of evidence, cost, values and preferences, and trade-offs between good and bad consequences. One challenge in using GRADE is learning the process itself, as systematic reviewers, policy makers, and TB experts are not necessarily trained in the GRADE approach. We expect this challenge to be overcome as more people receive training and use GRADE. Another challenge recognizes situations in which patient outcomes may not reflect the accuracy or benefit of a diagnostic test/approach because treatment is unavailable (e.g. improved microscopy in facilities where stock-outs of anti-TB drugs occur frequently). Additional limitations and challenges for diagnostic policies are summarized in Table 5. A recent review by Kavanagh [66] provides an interesting perspective on GRADE, especially on the issue of whether GRADE itself is reliable and has been proven to be valid.

By the nature of the GRADE process being based on evidence, it is intrinsically reliant on the availability and quality of the evidence base itself. As we have discussed above, challenges remain to ensure both the quality of primary diagnostic evaluations and the availability of the necessary types of data in systematic reviews. This is brought into clear focus when using the GRADE process, as a lack of objective studies on a topic opens the door to the substitution of expert opinion for evidence. Although expert experiences cannot be discounted, they may often not be generalizable and are subject to being influenced by personal agendas and anecdotal experiences. Experts in TB often rate the same evidence inconsistently, depending on their prior experience with a test, and this can result in poor interrater agreement on GRADE elements. For example, TB researchers who work extensively in resource-poor settings are often skeptical of high-tech tools and tend to undervalue them because of the perceived limited applicability in developing countries.

Conflicts of interest (COI) among guideline panel members and industry involvement in guideline processes are other issues of concern, especially when commercial tests and products are involved. There is some evidence that industry involvement is fairly common with TB diagnostic research, with about 40–50% of TB diagnostic studies reporting some degree of industry involvement or support [26,39]. A recent survey of IGRA guidelines and statements from various countries found that only a small minority had explicit COI disclosures [67]. Some organizations have recognized the need to address the issue of COI. For example, the American Thoracic Society (ATS) published its COI policy for guideline development in 2009 [68]. This policy now recommends procedures such as self-declaration of COI; review of potential participants' COI; disclosure of COI to project participants; refusal or excusal from certain decisions or recommendations when appropriate; and disclosure of COI to users of documents or attendees of conferences. All agencies and bodies involved in guideline development should follow this example.

COI, however, are not restricted to commercial products. Diagnostic tests developers can be academics with no industry involvement. Because of their heavy intellectual investment in new test development and better understanding of the test, they tend to have strong opinions on how policies should be formulated and this can pose conflicts during the guideline development process. Should test developers be included in guideline panels, but excused from voting on recommendations? Sometimes, test developers publish systematic and narrative reviews on their own tests (which invariably tend to be positive) and it is unclear whether such reviews should be included or excluded in the GRADE process. Publication bias is an added concern, especially if industry-supported diagnostic studies are more likely to be published when they report positive findings. Unlike RCTs, inclusion of unpublished diagnostic studies is difficult because of the lack of a diagnostic trials registry.

The involvement of public–private partnerships for product development perhaps increases the complexity. These are often characterized by a partnership between a nonprofit organization and a for-profit diagnostics company with confidential agreements on intellectual property related to a co-developed diagnostic. Test developers from the nonprofit organization may have the same intellectual investment COI as test developers in academia, but may in addition have a COI related to their partnership with a for-profit company. These issues point out a fundamental problem with all guidelines, a problem that GRADE can never address – the fate of a guideline or policy can heavily rest on the group of experts and stakeholders included in the guideline development committee or panel.

The application of the GRADE approach to evidence on diagnostics is relatively new and as a result there are some

Table 5 Challenges and limitations in formulating tuberculosis diagnostic policies

Challenge or limitation	Description and examples
Limitations of the existing evidence base	Majority of TB diagnostic studies are focused on test accuracy (sensitivity and specificity); therefore, systematic reviews are also focused on accuracy Test accuracy studies are often poorly designed, executed, and reported. Impact of tests on patient-important outcomes is rarely available. Accuracy studies are downgraded by GRADE for 'directness' and can never receive a rating of 'high-quality' evidence. Ease of implementation, resources required, cost-effectiveness, biosafety, and programmatic issues are critical for policy, but systematic reviews may not provide such data.
Evidence vs. expert opinion	Existing evidence does not meet the needs of policy makers. Outcomes that experts want and GRADE requires are often not available. In such situations, expert opinion tends to dominate and experts do not always agree; expert opinions are often based on their own unique experiences and anecdotes, which may not necessarily be generalizable or valid.
Difficulties in learning and using the GRADE system	Systematic reviewers, policy makers, and TB experts are not necessarily trained in GRADE. Grading may be done inconsistently across tests by different systematic reviewers; same evidence can be interpreted and rated differently; GRADE ratings may be revised *posthoc*, depending on which tests the experts want to recommend. Some tests are actively 'championed', whereas others are not and this can result in uneven decisions
Conflicts of interest and involvement of test developers	Participation of test developers and industry representatives in the policy process introduces conflicts of interest. There is no consensus on whether test developers and those invested in specific technologies be allowed to do systematic reviews and participate in guideline panel meetings. There can be tension between commercial and noncommercial tests; type and quality of evidence might differ for commercial vs. noncommercial tests, and commercial products might be more actively championed by those with industry involvement.
Patient-important outcomes	Patient outcomes may not reflect the accuracy or benefit of a diagnostic test/approach in settings with weak overall health infrastructure (e.g. rapid or improved microscopy in facilities where stock-outs of anti-TB drugs occur frequently). The possible tension (for TB diagnosis and control) between the importance of individual patient outcomes and public health outcomes (e.g. the notion that false-negative sputum smear results may pose a greater public health risk than false-positive results). For tests used at the central/reference laboratory level, patient-outcome data may not be a good index of a test's impact; the test's impact is confounded by several other factors such as specimen transport, time to get results back to the clinicians, weak healthcare systems, etc. Impact on patient outcomes is affected not just by the test, but the whole package, including treatment, healthcare system efficiency, etc. It can be difficult to separate out the test's impact, and hard/expensive to study the whole package or strategy (which can be time-consuming and expensive). Diagnostic RCTs are rarely available and very hard to do (ethics, cost, etc.) In addition to patient values and preferences, need to acknowledge preferences and values of laboratory technologists and test users. If RCTs and patient-important outcomes are required for noncommercial tests, this will be severely limited by access to funds required to perform these large-scale evaluations.
Systematic review methods	No standardized methodology to search for and objectively synthesize evidence on operational implementation issues, costs to health services, costs to patients, and patient perspectives on new diagnostic tests and approaches. Narrative evidence on the above issues may be excluded from search strategies during systematic reviews of studies on diagnostic accuracy. Results from qualitative and socio-economic studies may not have been captured in the systematic reviews on diagnostic accuracy of the different approaches. Systematic reviews can make an effort to look for, include, and describe outcomes other than sensitivity and specificity, but often do not because they choose to focus instead on easily meta-analyzable outcomes. Policy makers should have a thorough understanding of all the important outcomes (including outcomes that are important to patients) they hope to include in their policy deliberations before commissioning systematic reviews. By explicitly outlining the test characteristics that will influence their decisions in advance, guideline panels can ensure evidence is as complete and objective as possible. This approach will minimize evidence gaps, making the process less susceptible to expert opinion. Weighting the importance of test characteristics in advance can also help to avoid redefining and reinterpreting evidence *posthoc* to suit individual desires to recommend or not recommend.

GRADE, Grading of Recommendations Assessment, Development and Evaluation; RCT, randomized controlled trial; TB, tuberculosis.

difficulties specific to diagnostics, which may be alleviated in time. For example, forcing diagnostic evidence into the RCT framework can be nonintuitive to laboratory researchers who typically conduct diagnostic evaluations. Certainly, the lack of experience using GRADE on the part of systematic reviewers and policy makers currently can lead to inconsistent interpretation of criteria and the revision of ratings *posthoc* in order to create GRADE profiles consistent with predetermined opinions regarding diagnostics that should be recommended. The transition from traditional policy making, which was made primarily based on expert opinion, to the use of more standardized, objective methods is likely to be a struggle for all organizations whether it is clearly acknowledged and dealt with or not.

The absence of diagnostic RCTs and data regarding patient-important outcomes and preferences in the field of TB diagnostics is a major hindrance to their assessments using GRADE, which currently places much weight on these aspects of patient care. As noted above, studies providing estimates of accuracy alone are downgraded for their lack of 'direct' evidence and thus cannot achieve a rating of 'high-quality' evidence. Although it can be agreed that higher levels of evidence need to be encouraged when assessing diagnostics, there are many practical barriers to extrapolating between the use of a diagnostic and the clinical outcomes of patients. Any number of deficiencies in the health system can impact a patient outcome, some of which may prevent the full recognition of benefits clearly provided by a diagnostic. At the same time, many user-important outcomes (as described above), which are of great importance to the feasibility of implementing diagnostics, are not easily captured in the GRADE process.

Diagnostic RCTs are almost nonexistent in TB. Even if they were feasible, there are concerns about their design, interpretation, and ethics [69]. Diagnostic RCTs do not just evaluate a test; they evaluate a strategy or package that includes testing followed by some intervention as a follow-up to the test result [44]. In this context, it is not easy to disentangle the efficacy of the test from the efficacy of the follow-up treatment or intervention. Furthermore, it is not easy to capture patient-important outcomes when ethical considerations prevent clinical decision-making on the basis of a trial product. Evidence from RCTs in highly controlled trial settings may not reflect the real-world conditions in which diagnostics have to be ultimately deployed. Lastly, diagnostic RCTs can take much longer than conventional diagnostic accuracy studies and this can delay the introduction of new policies.

The lack of stringent regulation and licensing of diagnostics certainly contributes to the lack of standardized, high-quality evidence available for the use of decision and policy makers. Additionally, this leads to the need for diagnostic policy processes to not only assess 'added benefit' of one test over another, but often to make the first objective assessment of a test's performance. The imposition of well defined, high standards at the stage of regulatory approval would help guide developers and researchers in their assessments of new diagnostics and provide impetus for the publication of appropriate and needed evidence. Compared to the therapeutics arena wherein strict regulation is imposed before a product is licensed for use, diagnostics require very limited data before they can be used to make patient care decisions. For example, despite a large body of evidence showing poor accuracy of commercial serological, antibody detection tests for TB, several commercial serological tests are on the market and used frequently in developing countries with weak regulatory systems [16,17,70,71]. Poorly performing diagnostics continue to remain on the market despite poor performance in the published literature and there are no mechanisms to 'withdraw' or 'ban' a bad diagnostic.

It needs to be recognized that by the nature of systematic reviews (upon which the GRADE process is reliant), the questions which are asked are of paramount importance [72]. Search criteria, selection processes, and presentation of evidence will all depend on the exact questions posed. If policy makers have a clear understanding of the issues that are important for implementation of a given diagnostic in advance, then evidence can be objectively collected to inform decisions and assessments on both quantitative and qualitative aspects. However, if only issues of test accuracy and technical performance are covered by systematic reviews, then gaps pertaining to other aspects of performance may need to be filled through less objective expert opinion.

All things considered, policy making is a big challenge in TB, as it is in other areas of medicine. Although GRADE has its limitations and can definitely be improved and adapted for TB diagnostics, we believe it is a major advance over the conventional policy making process.

Challenges in translating policies into impact

Availability of new tools does not necessarily ensure their adoption and implementation. Translation of policy into practice requires better understanding of barriers to implementation and methods to overcome such barriers. The impact of new tests will depend largely on the extent of their introduction and acceptance into the global public sector. This will itself depend in part on policy decisions made by international technical agencies such as WHO, by donors, and ultimately by national TB programs. This area has been extensively reviewed by

the Stop TB Partnership's Task Force on Retooling and has led to the creation of a roadmap to guide global, regional, and country-based activities as well as guidelines for engaging stakeholders in retooling and the introduction of specific TB diagnostics [73–75]. The work of the time-limited and now disbanded Task Force on Retooling has been mainstreamed into routine TB control activities led by the DOTS Expansion Working Group and its Subgroup on Introducing New Tools and Approaches (INAT).

The major obstacles to diagnostic retooling for TB control are undoubtedly the poor laboratory infrastructure and weak healthcare delivery systems present in many disease-endemic countries [76]. This has been recognized for many years. Although vastly increased funds are being invested in diagnostics retooling through national investments and funding agencies, there is still little guidance available to countries on what new diagnostic tools, or combinations of these tools, should be implemented in their particular epidemiological/health systems settings, what laboratory capability or capacity should be built to support this implementation, or how this should be done. A roadmap for strengthening TB laboratories that is abreast with recent developments and addresses these issues is urgently needed [77]. Beyond introducing new diagnostics and strengthening laboratories, challenges will remain in the development of accessible, equitable, and high-quality diagnostic services based on them and ensuring that healthcare delivery systems are strengthened so that better diagnostic services translate into better care [78]. In many countries, the private healthcare sector is the dominant source of healthcare. Lack of private sector involvement in TB control is a major weakness in existing programs.

Conclusion

After decades of neglect and poor progress, there is now great excitement about the development and introduction of new diagnostics for TB. The diagnostics pipeline has rapidly expanded and several new tools and strategies have received WHO endorsement for implementation at the country level. There are major gaps in the existing pipeline and the evidence base is predominantly made up of research studies of test accuracy. Future TB diagnostic research needs to focus on clinically meaningful outcomes and also consider obstacles to implementation. The GRADE system has brought greater transparency and evidence-based approaches to policy making, though GRADE for diagnostics is still a work in progress. Future TB policies and guidelines will need to be transparent, evidence-based, and free of COI. Today, despite many years of intensive effort to remedy the situation, weak laboratories remain the major immediate obstacle to translating policy into practice in low-income and middle-income countries. With the engagement of all key stakeholders, these challenges can be addressed to translate all the scientific progress into public health impact.

Acknowledgements

M.P. is a recipient of a New Investigator Award from the Canadian Institutes of Health Research (CIHR). J.M. is a recipient of a Quebec Respiratory Health Training Program (QRHTP) Fellowship. These funding sources had no role in the preparation of this manuscript, nor the decision to submit the manuscript for publication.

The authors have no financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. M.P. serves as an external consultant for the Foundation for Innovative New Diagnostics (FIND). FIND is a nonprofit agency that works with several industry partners in developing and evaluating new diagnostics for neglected infectious diseases. M.P. serves as a co-chair of the Stop TB Partnership's New Diagnostics Working Group (NDWG) and A.R. serves as the secretary of the NDWG. K.S. serves as co-chair of the Evidence Synthesis subgroup of the NDWG.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

Additional references related to this topic can also be found in the Current World Literature section in this issue (pp. 286–287).

1 • Pai M, Minion J, Sohn H, *et al.* Novel and improved technologies for tuberculosis diagnosis: progress and challenges. Clin Chest Med 2009; 30:701–716.
This paper reviews the existing evidence base on TB diagnostics, describes the progress of new technologies, and ends with a review of cost-effectiveness and modeling studies on the potential effect of new diagnostics in TB control.

2 Perkins MD, Cunningham J. Facing the crisis: improving the diagnosis of tuberculosis in the HIV era. J Infect Dis 2007; 196 (Suppl 1):S15–S27.

3 • World Health Organization & Stop TB Partnership. New laboratory diagnostic tools for tuberculosis control. Geneva: World Health Organization; 2008.
Detailed review of the existing TB diagnostics pipeline.

4 Pai M, O'Brien R. New diagnostics for latent and active tuberculosis: state of the art and future prospects. Semin Respir Crit Care Med 2008; 29:560–568.

5 • Pai M, Ramsay A, O'Brien R. Comprehensive new resource for evidence-based TB diagnosis. Expert Rev Mol Diagn 2009; 9:637–639.
This study describes the development of a new website resource on 'evidence-based TB diagnosis'.

6 • Pai M, Ramsay A, O'Brien R. Evidence-based tuberculosis diagnosis. PLoS Med 2008; 5:e156.
A survey of all the systematic reviews on TB diagnostic tests and the contribution made by systematic reviews for informing clinical practice and policy.

7 Steingart KR, Henry M, Ng V, *et al.* Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis 2006; 6:570–581.

8 Steingart KR, Ng V, Henry M, *et al.* Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis 2006; 6:664–674.

9 Mase SR, Ramsay A, Ng V, *et al.* Yield of serial sputum specimen examinations in the diagnosis of pulmonary tuberculosis: a systematic review. Int J Tuberc Lung Dis 2007; 11:485–495.

10 Greco S, Girardi E, Navarra S, Saltini C. The current evidence on diagnostic accuracy of commercial based nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis. Thorax 2006; 61:783–790.

11 Ling DI, Flores LL, Riley LW, Pai M. Commercial nucleic-acid amplification tests for diagnosis of pulmonary tuberculosis in respiratory specimens: meta-analysis and meta-regression. PLoS One 2008; 3:e1536.

12 Dinnes J, Deeks J, Kunst H, *et al.* A systematic review of rapid diagnostic tests for the detection of tuberculosis infection. Health Technol Assess 2007; 11:1–196.

13 Pai M, Flores LL, Hubbard A, *et al.* Nucleic acid amplification tests in the diagnosis of tuberculous pleuritis: a systematic review and meta-analysis. BMC Infect Dis 2004; 4:6.

14 Pai M, Flores LL, Pai N, *et al.* Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. Lancet Infect Dis 2003; 3:633–643.

15 Daley P, Thomas S, Pai M. Nucleic acid amplification tests for the diagnosis of tuberculous lymphadenitis: a systematic review. Int J Tuberc Lung Dis 2007; 11:1166–1176.

16 Steingart KR, Henry M, Laal S, *et al.* A systematic review of commercial serological antibody detection tests for the diagnosis of extra-pulmonary tuberculosis. Thorax 2007; 62:911–918.

17 Steingart KR, Henry M, Laal S, *et al.* Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: a systematic review. PLoS Med 2007; 4:e202.

18 Steingart KR, Dendukuri N, Henry M, *et al.* Performance of purified antigens for serodiagnosis of pulmonary tuberculosis: a meta-analysis. Clin Vaccine Immunol 2009; 16:260–276.

19 Greco S, Girardi E, Masciangelo R, *et al.* Adenosine deaminase and interferon gamma measurements for the diagnosis of tuberculous pleurisy: a meta-analysis. Int J Tuberc Lung Dis 2003; 7:777–786.

20 Liang QL, Shi HZ, Wang K, *et al.* Diagnostic accuracy of adenosine deaminase in tuberculous pleurisy: a meta-analysis. Respir Med 2008; 102:744–754.

21 Jiang J, Shi HZ, Liang QL. Diagnostic value of interferon-g in tuberculous pleurisy: a meta-analysis. Chest 2007; 131:1133–1141.

22 Kalantri S, Pai M, Pascopella L, *et al.* Bacteriophage-based tests for the detection of Mycobacterium tuberculosis in clinical specimens: a systematic review and meta-analysis. BMC Infect Dis 2005; 5:59.

23 Cruciani M, Scarparo C, Malena M, *et al.* Meta-analysis of BACTEC MGIT 960 and BACTEC 460 TB, with or without solid media, for detection of mycobacteria. J Clin Microbiol 2004; 42:2321–2325.

24 Farhat M, Greenaway C, Pai M, Menzies D. False-positive tuberculin skin tests: what is the absolute effect of BCG and nontuberculous mycobacteria? Int J Tuberc Lung Dis 2006; 10:1192–1204.

25 Wang L, Turner MO, Elwood RK, *et al.* A meta-analysis of the effect of Bacille Calmette Guerin vaccination on tuberculin skin test measurements. Thorax 2002; 57:804–809.

26 Pai M, Zwerling A, Menzies D. T-cell based assays for the diagnosis of latent tuberculosis infection: an update. Ann Intern Med 2008; 149:177–184.

27 Menzies D, Pai M, Comstock G. Meta-analysis: new tests for the diagnosis of latent tuberculosis infection: areas of uncertainty and recommendations for research. Ann Intern Med 2007; 146:340–354.

28 Pai M, Riley LW, Colford JM Jr. Interferon-gamma assays in the immunodiagnosis of tuberculosis: a systematic review. Lancet Infect Dis 2004; 4:761–776.

29 van Zyl-Smit RN, Zwerling A, Dheda K, Pai M. Within-subject variability of interferon-g assay results for tuberculosis and boosting effect of tuberculin skin testing: a systematic review. PLoS One 2009; 4:e8517.

30 Pai M, Kalantri S, Pascopella L, *et al.* Bacteriophage-based assays for the rapid detection of rifampicin resistance in Mycobacterium tuberculosis: a meta-analysis. J Infect 2005; 51:175–187.

31 Minion J, Pai M. Bacteriophage assays for rifampin resistance detection in Mycobacterium tuberculosis: updated meta-analysis. Presented at the World Health Organization Expert Group Meeting on Non-commercial Culture Methods and Mycobacteriophage-based Assays for Rapid Screening of Patients at Risk of Drug-Resistant Tuberculosis. September 8–9, 2009, Geneva, Switzerland.

32 Ling DI, Zwerling A, Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. Eur Respir J 2008; 32:1165–1174.

33 Morgan M, Kalantri S, Flores L, Pai M. A commercial line probe assay for the rapid detection of rifampicin resistance in Mycobacterium tuberculosis: a systematic review and meta-analysis. BMC Infect Dis 2005; 5:62.

34 Bwanga F, Hoffner S, Haile M, Joloba ML. Direct susceptibility testing for multi drug resistant tuberculosis: a meta-analysis. BMC Infect Dis 2009; 9:67.

35 Martin A, Portaels F, Palomino JC. Colorimetric redox-indicator methods for the rapid detection of multidrug resistance in Mycobacterium tuberculosis: a systematic review and meta-analysis. J Antimicrob Chemother 2007; 59:175–183.

36 Martin A, Panaiotov S, Portaels F, *et al.* The nitrate reductase assay for the rapid detection of isoniazid and rifampicin resistance in Mycobacterium tuberculosis: a systematic review and meta-analysis. J Antimicrob Chemother 2008; 62:56–64.

37 Minion J, Leung E, Menzies D, Pai M. Microscopic-Observation Drug Susceptibility (MODS) and Thin Layer Agar (TLA) assays for the detection of drug resistant tuberculosis: a systematic review. Presented at the World Health Organization Expert Group Meeting on Non-commercial Culture Methods and Mycobacteriophage-based Assays for Rapid Screening of Patients at Risk of Drug-Resistant Tuberculosis. September 8–9, 2009, Geneva, Switzerland.

38 Pai M, O'Brien R. Tuberculosis diagnostics trials: do they lack methodological rigor? Expert Rev Mol Diagn 2006; 6:509–514.

39 Fontela PS, Pai NP, Schiller I, *et al.* Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One 2009; 4:e7753.

40 Banoo S, Bell D, Bossuyt P, *et al.* Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol 2006; 4 (9 Suppl):S21–S31.

41 Bossuyt PM, Reitsma JB, Bruns DE, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003; 138:40–44.

42 •• Stop TB Partnership's New Diagnostics Working Group and World Health Organization. Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics. Geneva: World Health Organization; 2009.

Comprehensive, well referenced blueprint to guide researchers, clinicians, industry partners, academics, and TB controllers in all sectors in all aspects of TB diagnostics development, from concept to evaluation, implementation, scale-up, delivery, and impact.

43 Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009; 29:E30–E38.

44 Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009; 29:E1–E12.

45 Keeler E, Perkins MD, Small P, *et al.* Reducing the global burden of tuberculosis: the contribution of improved diagnostics. Nature 2006; 444 (Suppl 1): 49–57.

46 • Abu-Raddad LJ, Sabatelli L, Achterberg JT, *et al.* Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. Proc Natl Acad Sci U S A 2009; 106:13980–13985.

Recent modeling study on potential impact and epidemiological benefits of new TB diagnostic tools.

47 Getahun H, Harrington M, O'Brien R, Nunn P. Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings: informing urgent policy changes. Lancet 2007; 369: 2042–2049.

48 Marais BJ, Pai M. New approaches and emerging technologies in the diagnosis of childhood tuberculosis. Paediatr Respir Rev 2007; 8:124–133.

49 Barry CE 3rd, Boshoff HI, Dartois V, *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nat Rev Microbiol 2009; 7:845–855.

50 Pai M. Spectrum of latent tuberculosis: existing tests cannot resolve the underlying phenotypes. Nat Rev Microbiol 2010 Jan 19 [Epub ahead of print].

51 • World Health Organization. Moving research findings into new WHO policies. http://www.who.int/tb/dots/laboratory/policy/en/index4.html. Geneva: World Health Organization; 2008.

This WHO document describes the new process for policies on tuberculosis.

52 • Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149:889–897.

This study describes the Cochrane Collaboration methodologies and approaches for diagnostic test accuracy meta-analyses.

53 World Health Organization. The use of liquid medium for culture and DST. http://www.who.int/tb/dots/laboratory/policy/en/index3.html. Geneva: World Health Organization; 2007.

54 World Health Organization. Reduction of number of smears for the diagnosis of pulmonary TB [cited 27 January 2010]. http://www.who.int/tb/dots/laboratory/policy/en/index2.html

55 World Health Organization. Definition of a new sputum smear-positive TB case [cited 27 January 2010]. http://www.who.int/tb/dots/laboratory/policy/en/index1.html

56 World Health Organization. Policy statement. Molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis (MDR-TB) [cited 27 January 2010]. http://www.who.int/tb/dots/laboratory/policy/en/index4.html

57 World Health Organization. Report of the 9th Meeting of the Strategic and Technical Advisory Group on Tuberculosis (STAG-TB). Geneva: World Health Organization; 2009 http://www.who.int/tb/advisory_bodies/stag/en/index.html

58 Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on •• rating quality of evidence and strength of recommendations. BMJ 2008; 336:924–926.
This study provides a comprehensive introduction to the GRADE approach for developing policy recommendations.

59 Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. Lancet 2007; 369:1883–1889.

60 Schunemann HJ, Oxman AD, Brozek J, *et al.* GRADE: assessing the quality of evidence for diagnostic recommendations. Evid Based Med 2008; 13:162–163.

61 World Health Organization. Guidelines for WHO Guidelines. Geneva: World Health Organization; 2003.

62 World Health Organization. WHO policy on TB infection control in health-care facilities, congregate settings and households. Geneva: World Health Organization; 2010.

63 World Health Organization. Treatment of tuberculosis. 4th ed. Geneva: World Health Organization; 2009.

64 Schunemann HJ, Oxman AD, Brozek J, *et al.* Grading quality of evidence and •• strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336:1106–1110.
This report outlines the GRADE approach to diagnostic tests and strategies.

65 Brozek JL, Akl EA, Jaeschke R, *et al.* Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009; 64:1109–1116.

66 Kavanagh BP. The GRADE system for rating clinical guidelines. PLoS Med 2009; 6:e1000094.

67 Pai M. Guidelines on IGRAs: concordant or discordant? 2nd Global Symposium on IGRAs. Dubrovnik, Croatia; 2009.

68 Schunemann HJ, Osborne M, Moss J, *et al.* An official American Thoracic Society Policy statement: managing conflict of interest in professional societies. Am J Respir Crit Care Med 2009; 180:564–580.

69 Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. Ann Epidemiol 2006; 16:540–544.

70 Steingart KR, Ramsay A, Pai M. Commercial serological tests for the diagnosis of tuberculosis: do they work? Future Microbiol 2007; 2:355–359.

71 World Health Organization & UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases. Laboratory-based evaluation of 19 commercially-available rapid diagnostic tests for tuberculosis. Geneva: World Health Organization; 2008.

72 Pai M, McCulloch M, Enanoria W, Colford JM Jr. Systematic reviews of diagnostic test evaluations: what's behind the scenes? ACP J Club 2004; 141:A11–A13.

73 World Health Organization. New technologies for tuberculosis control: a framework for their adoption, introduction and implementation [WHO/HTM/STB/2007.40]. Geneva: World Health Organization; 2007.

74 Stop TB Partnership Retooling Task Force. Engaging stakeholders for retooling TB control. Geneva: World Health Organization; 2008.

75 Stop TB Partnership Retooling Task Force. Checklist of key actions for the use of liquid culture and drug susceptibility testing. Geneva: World Health Organization; 2008.

76 Ridderhof JC, van Deun A, Kam KM, *et al.* Roles of laboratories and laboratory systems in effective tuberculosis programmes. Bull World Health Organ 2007; 85:354–359.

77 World Health Organization. Strategic approach to the strengthening of laboratory services for tuberculosis control: 2006–2009. Geneva: World Health Organization; 2006.

78 Yagui M, Perales MT, Asencios L, *et al.* Timely diagnosis of MDR-TB under program conditions: is rapid drug susceptibility testing sufficient? Int J Tuberc Lung Dis 2006; 10:838–843.

INT J TUBERC LUNG DIS 14(12):1506–1507

EDITORIAL

Assessing the impact of new diagnostics on tuberculosis control

THE LAST FEW YEARS have seen an unprecedented effort to develop new diagnostics for tuberculosis (TB), and a number of significant achievements in evidence-based TB diagnosis and translation of evidence into policy.[1–3] The World Health Organization (WHO) has endorsed more than 10 new or improved tools for the diagnosis of TB since 2007, and four additional tools are currently under review.[1,2]

The WHO process for policy formulation related to TB is a five-step process, consisting of: 1) identifying the need for a policy change; 2) reviewing the evidence; 3) convening an expert panel; 4) assessing draft policies and guidelines; and 5) formulating and disseminating policy.[4]

Central to this policy-making process is the synthesis of the available evidence on a diagnostic through systematic reviews and the application of the GRADE (grading of recommendations assessment, development and evaluation) approach to guideline development.[5] The GRADE approach often rates diagnostic studies reporting only test accuracy (e.g., sensitivity, specificity) as low-quality evidence for policy development because the link between diagnostic accuracy and patient-important outcomes is indirect.[5] A large proportion (>80%) of all TB diagnostic research publications are focused on test accuracy, and there is little published evidence on the impact of TB tests on patient-important outcomes.[2,6] To fill this gap, expert opinion is sought, through WHO Expert Group Meetings, on the likely effect of diagnostic accuracy on patient-important outcomes for a given diagnostic. As a result, strong recommendations (positive or negative) have been made on the basis of moderate or low-quality evidence,[7] and the GRADE approach permits this.[5]

There is a need to conduct diagnostics evaluations that assess the impact of new diagnostics on patient-important outcomes, including time to diagnosis, time to treatment, incremental value of new diagnostics, impact of new tests on clinician decision making, appropriateness of the treatment regimen offered on the basis of the diagnostic test result and impact of testing on treatment outcomes. While the methodologies for evidence synthesis and policy recommendation have improved greatly over recent years and the value of these activities is being recognised, most of the original research being synthesised reports only diagnostic accuracy, and policy recommendations continue to be made on the basis of moderate/low-quality evidence.[2,3]

There are recognised challenges in assessing a number of major patient-important outcomes through a diagnostic trial.[2] Foremost among these are the ethical considerations that would prevent patient-management decisions being based upon the result of the trial diagnostic. It is also recognised that most TB diagnostics will not be used alone, but in combination with clinical decisions to test, and with other diagnostics such as smear microscopy and/or culture and drug susceptibility testing. This complexity is likely to influence the patient-important outcomes. It is thus the patient-important outcomes associated with use of a particular diagnostic-intervention package that are the outcomes of interest.

The Stop TB Partnership's New Diagnostics Working Group (NDWG) has recently published a scientific blueprint for TB diagnostics development and evaluation.[8] This blueprint presents an overview of what evidence is required for the comprehensive assessment of a diagnostic. This document makes it clear that it is not only *patient*-important outcomes that need to be measured, but also *population*-important outcomes (e.g., gender equity in access to the diagnostic) and *health systems*-important outcomes (e.g., demands of new diagnostics on human resources). Most diagnostic accuracy studies, as well as the so-called 'demonstration studies', are conducted in controlled settings with technical and financial inputs far in excess of the resources available in routine programme conditions. Valid measures of some population- and health systems-important outcomes will be difficult to obtain until the tool has been introduced into routine National Tuberculosis Programme (NTP) activities. Beyond populations and health systems, it is necessary to conduct research on the public health and epidemiological impact of introducing new diagnostics to ensure that the intervention is associated with improved case detection and cure rates and reduced TB transmission and incidence. The term 'impact' thus has different interpretations, ranging from the impact of a test on an individual patient's outcome to the epidemiological impact of widespread scale-up of a new diagnostic test in a population.

In a perfect world, we would want all the evidence available before making policy recommendations. However, collecting such a body of evidence will be time-consuming and expensive. Waiting for all the evidence could delay, by several years, the uptake of a new tool with the potential to dramatically improve TB control. The NDWG's scientific blueprint recognises that research is needed both before and after WHO endorsement and introduction of a new tool into NTP activities. Despite the WHO endorsement of several new diagnostic tools in recent years, we are unaware of any peer-reviewed publications describing the impact of new TB diagnostics on patient-, population- or

[A version in French of this article is available from the Editorial Office in Paris and from the Union website www.theunion.org]

health systems-important outcomes or on the epidemiological impact. Modelling studies on the likely impact of new diagnostics have been published,[9,10] but they will need to be supported by real-world empiric data on impact. Such studies would stimulate the scale-up of new diagnostics (if associated with positive impact) or permit the revision or amendment of WHO recommendations (if associated with negative or no impact or shown not to be cost-effective compared to alternative approaches). At present, the five-step policy making process does not include a formal stage in which the WHO reviews its recommendations in the light of experience from routine NTP practice and other impact studies.

A major obstacle to conducting impact assessments is a lack of consensus on what 'impact' really means, and what patient-, population-, health systems-, and epidemiology-important outcomes should be measured to decide on impact. There is also a lack of guidance on the methods (i.e., study designs) to be used to measure them, which methods will be most rapid and cost-effective, and who exactly should be assigned the responsibility of measuring impact (academia, industry, product development partnerships, NTPs, the WHO, technical agencies or other independent bodies).

The Impact Assessment Framework (IAF) described by Mann and colleagues in this issue of the *Journal*[11] is a welcome step in dealing with this gap in our knowledge and practice. Suggesting an initial definition of 'impact', the authors present, for the first time, a systematic, multi-layered approach to collecting relevant data on the overall impact of introducing new diagnostic technologies for TB. A number of methods and study designs exist that could be used to collect data in the different layers proposed in the IAF. However, some kind of consensus and guidance is necessary on what outcomes to measure and how to use the different methodologies to comprehensively collect the evidence required. Efforts are underway to convene an expert group to provide concrete guidance on assessing the impact of new diagnostics on TB control, especially for WHO-approved tests.

Although the methods and approaches for systematic reviews of diagnostic accuracy have improved greatly in the last few years,[12] those used to synthesise the data from studies on equity in health services, for example, are much less developed. In anticipation of new types of data becoming available for policy guidance, evidence synthesis experts need to develop the appropriate tools for their review, and the GRADE approach will need to evolve based on accumulated experience.

It is critical that donors and institutions supporting TB control be aware of the need to continue research within NTP activities beyond WHO endorsement of a new diagnostic tool. Significant funding will be required for such implementation research. There also needs to be a formal mechanism by which the WHO reviews post-endorsement evidence on impact, and through which endorsements or recommendations on TB diagnostics can be revised, expanded or retracted.

Andrew Ramsay, PhD*
Karen R. Steingart, MD, MPH†
Madhukar Pai, MD, PhD‡
**UNICEF/UNDP/World Bank/WHO*
Special Programme for
Research and Training in Tropical Diseases
World Health Organization
Geneva, Switzerland
†*Curry International Tuberculosis Center*
University of California, San Francisco
San Francisco, California, USA
‡*Department of Epidemiology,*
Biostatistics & Occupational Health
McGill University
Montreal, Quebec, Canada
e-mail: madhukar.pai@mcgill.ca

Acknowledgements

All the authors are involved with the Stop TB Partnership's New Diagnostics Working Group (NDWG). AR serves as the secretariat, KRS serves as co-chair of the evidence synthesis subgroup, and MP serves as co-chair of NDWG. MP is a consultant for the Bill & Melinda Gates Foundation. The views and opinions expressed by the authors are their own and do not necessarily reflect those of the Stop TB Partnership or the institutions for whom they work.

References

1 Wallis R S, Pai M, Menzies D, et al. Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. Lancet 2010; 375: 1920–1937.
2 Pai M, Minion J, Steingart K, Ramsay A. New and improved tuberculosis diagnostics: evidence, policy, practice, and impact. Curr Opin Pulm Med 2010; 16: 271–284.
3 Pai M, Ramsay A, O'Brien R. Evidence-based tuberculosis diagnosis. PLoS Med 2008; 5: e156.
4 World Health Organization. Moving research findings into new WHO policies. Geneva, Switzerland: WHO, 2008. http://www.who.int/tb/dots/laboratory/policy/en/index4.html Accessed September 2010.
5 Schunemann H J, Oxman A D, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336: 1106–1110.
6 Brunet L, Minion J, Lienhardt C, Pai M. Mapping the landscape of tuberculosis diagnostic research. Am J Respir Crit Care Med 2010; 181: A2255.
7 WHO Expert Group meeting reports on sputum smear microscopy and non-commercial culture and bacteriophage-based assays. Geneva, Switzerland: WHO, 2009. http://www.who.int/tb/dots/laboratory/policy_statements/en/index.html Accessed October 2010.
8 Stop TB Partnership's New Diagnostics Working Group/World Health Organization. Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics. Geneva, Switzerland: World Health Organization, 2009.
9 Abu-Raddad L J, Sabatelli L, Achterberg J T, et al. Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. Proc Natl Acad Sci USA 2009; 106: 13980–13985.
10 Dowdy D W, Chaisson R E, Maartens G, Corbett E L, Dorman S E. Impact of enhanced tuberculosis diagnosis in South Africa: a mathematical model of expanded culture and drug susceptibility testing. Proc Natl Acad Sci USA 2008; 105: 11293–11298.
11 Mann G, Squire S B, Bissell K, et al. Beyond accuracy: creating a comprehensive evidence base for TB diagnostic tools. Int J Tuberc Lung Dis 2010; 14: 1518–1524.
12 Leeflang M M, Deeks J J, Gatsonis C, Bossuyt P M. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149: 889–897.

INT J TUBERC LUNG DIS 14(12): 1518–1524

STATE OF THE ART

Beyond accuracy: creating a comprehensive evidence base for tuberculosis diagnostic tools

G. Mann,* S. B. Squire,* K. Bissell,† P. Eliseev,‡ E. Du Toit,§ A. Hesseling,§ M. Nicol,¶ A. Detjen,† A. Kritski#

***Liverpool School of Tropical Medicine, Liverpool, UK; †International Union Against Tuberculosis and Lung Disease, Paris, France; ‡Arkhangelsk Regional Anti-tuberculosis Dispensary, Arkhangelsk, Russian Federation; §Desmond Tutu TB Centre, Cape Town, ¶National Health Laboratory Services, Johannesburg, South Africa; #Rede TB, Brasília, Brazil**

SUMMARY

The need for a strong and comprehensive evidence base to support decision making with regard to the implementation of new and improved diagnostic tools and approaches has been highlighted by a number of stakeholders; these include members of the New Diagnostics Working Group (NDWG) and the Subgroup for Introducing New Approaches and Tools of the Stop TB Partnership. To compile such evidence in a systematic manner, we have developed an impact assessment framework (IAF) which links evidence on inputs to outcomes.

The IAF comprises five interconnected layers: effectiveness analysis, equity analysis, health systems analysis, scale-up analysis and policy analysis. It can be used by new diagnostics developers and other interested research teams to collect as much policy-relevant data as possible prior to, during and after the demonstration phase of tool development. The evidence collated may be used by international and national policy makers to support adoption, implementation and scale-up decisions. The TREAT TB (Technology, Research, Education and Technical Assistance for TB) initiative uses the IAF in its operational research and field evaluations of new tools and approaches for TB diagnosis. It has also been incorporated into the NDWG's recent publication: 'Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics'. This article describes the IAF and the process of improving it and suggests next steps in overcoming the challenges in its implementation.

KEY WORDS: impact; evidence; tuberculosis; policy; diagnostics

Every TB patient must have access to an effective diagnosis, treatment and cure.
—The Global Plan to Stop TB 2006–2015[1]

THE ABOVE PRINCIPLE is key to attaining the vision of the Stop TB Partnership of seeing a tuberculosis (TB) free world by 2050.[1] Access to an 'effective diagnosis' has long been a concern. Smear microscopy is not sufficiently sensitive to detect tuberculosis disease in many cases, and particularly not in children and those who are co-infected with the human immunodeficiency virus. Multi- and extensively drug-resistant TB present new diagnostic and treatment challenges. Thankfully, new diagnostic approaches using existing tools have been recommended and new tools are in the pipeline.[2]

Any new approach or tool must be evaluated before being adopted by national tuberculosis programmes (NTPs); huge sums of money are already spent on TB diagnostics—estimated at more than $1 billion per year globally[3]—and resource-poor countries cannot afford to invest in interventions that are not more cost-effective than those already available. The World Health Organization (WHO) plays a key role in approving and developing guidelines for the use of new tools. The policy making process, described in a recent WHO statement,[4] comprises:

1 identifying the need for a policy change (e.g., the emergence of a new technology);
2 reviewing the evidence (e.g., through commissioning systematic reviews);
3 convening an expert panel to review evidence using the GRADE approach (Grading of Recommendations Assessment, Development and Evaluation, see BMJ 2008[5–9]);
4 assessing draft policies and guidelines (through the Strategic and Technical Advisory Group for TB, STAG-TB); and
5 formulating and disseminating new policies and guidelines.

Recent papers by Pai et al. have noted that systematic reviews, and hence the evidence reviewed in the above policy development process, have concentrated

Correspondence to: Gillian Hazel Mann, Liverpool School of Tropical Medicine–CRESTHA, Pembroke Place Liverpool L3 5QA, UK. Tel: (+44) 151 705 3139. Fax: (+44) 151 705 3743. e-mail: ghmann@liv.ac.uk

[A version in French of this article is available from the Editorial Office in Paris and from the Union website www.theunion.org]

mainly on test accuracy.[10,11] While such data are necessary, they are not sufficient to assess the contribution new diagnostics can make to the universal access requirements outlined in the Global Plan to Stop TB. A number of stakeholders, the Subgroup for Introducing New Approaches and Tools (INAT) and the New Diagnostics Working Group (NDWG) of the Stop TB Partnership, among others,[11–13] have called for a strong and comprehensive evidence base to support decision making with regard to implementation of new and improved diagnostic tools.

The NDWG has published 'Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics',[2] which outlines the required phases from needs assessment through test development to assessment of epidemiological impact, and all stages in between (Figure). The Stop TB Partnership's Retooling Taskforce, a precursor to INAT, stipulated the need for evidence that captures not only the benefits of new tools but also the risks and health systems implications associated with them.[14] This range of evidence is encapsulated in the Organisation for Economic Cooperation and Development definition of 'impact' subscribed to by multi- and bilateral donors who have signed the Paris Declaration on Aid Effectiveness (2005), which states that impact consists of:

> [The] positive and negative long-term effects on identifiable population groups produced by a development intervention, directly or indirectly, intended or unintended. These effects can be economic, socio-cultural, institutional, environmental, technological or of other types.[15]

Thus, measuring the impact of a diagnostic tool or approach involves assessing its positive and negative effects on different stakeholders (patients, health systems, laboratories, etc).

This entails summarising evidence not only about the test's accuracy, but also its effectiveness in field conditions in terms of diagnosing patients with various TB presentations, especially for the most infectious, and ensuring that they start appropriate treatment, its affordability, ease of implementation and potential for scale-up (for the health system) and accessibility (especially to poor and vulnerable TB suspects). Articulating and communicating this overall impact succinctly and with sufficient evidence is essential. It is required for national policy makers to make rational decisions about which new diagnostic tests and approaches to adopt, when and how to implement them, how to manage and finance them and how to ensure sustainable access and appropriate use.[14]

To compile such evidence in a systematic manner, we have developed an impact assessment framework (IAF) that links evidence on inputs to outcomes. This framework has been included in the NWDG's blueprint and has been adopted by the TREAT TB (Technology, Research, Education and Technical Assistance for TB) initiative for use in its operational research and field evaluations of new tools and approaches that are at late stages of development or have recently achieved international approval for use in TB diagnosis and treatment.

The present study describes the IAF, provides examples of how it can be used and suggests means of overcoming the remaining challenges in its implementation.

THE IMPACT ASSESSMENT FRAMEWORK

The IAF has been developed by a multidisciplinary team at the Liverpool School of Tropical Medicine and collaborators, including clinicians, laboratory specialists, health economists, social scientists and health systems analysts. It is based on a range of prior research activities in different countries that supported different elements of the evidence base.[16–25] These elements have been combined to provide an overarching

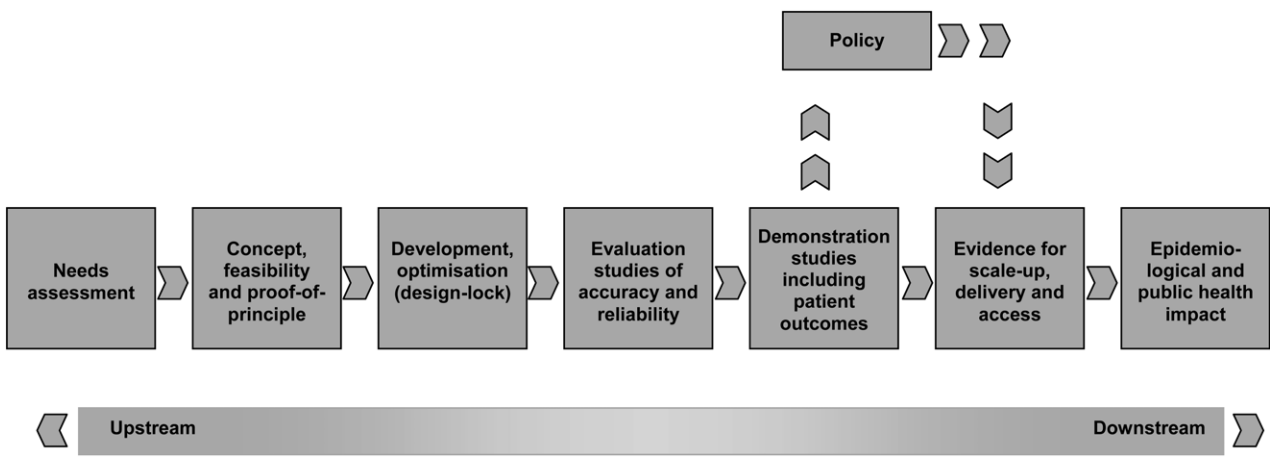


Figure Pathway for the development of tuberculosis diagnostics, from needs assessment to delivery [reproduced from reference 2 with permission from the World Health Organization].

framework (the IAF) to indicate how sufficient information for policy decisions could be collected in a systematic manner for all new diagnostic tools and approaches. The sufficiency of information has been considered in line with the international targets of the Global Plan to Stop TB and the Millennium Development Goals (MDGs).[26] The IAF, with references relating to different types of evidence, is shown in Table 1.

The IAF comprises five interconnected layers:

1 Layer 1: Effectiveness analysis
2 Layer 2: Equity analysis
3 Layer 3: Health systems analysis
4 Layer 4: Scale-up analysis
5 Layer 5: Policy analysis

Table 1 The impact assessment framework

Layer of assessment	Kinds of question(s) being addressed	References to studies addressing these questions
Layer 1 Effectiveness analysis	How well does the new tool work in terms of accuracy?	16
	How many additional cases will be identified who would otherwise not have been identified?	20
	How many additional cases will actually start (and complete) treatment as a result of using the new tool?	21
Layer 2 Equity analysis	Who benefits from the new tool (ambulant vs. hospitalised, poor/less poor, men/women, adults/children)?	27
	Why do these benefits accrue (level health system in which new diagnostic is deployed, change in time to issue of results, change in patient costs)?	22
Layer 3 Health system analysis	What are the human resource implications of introducing the new tool (training, number and cadre of staff)?	19
	What are the infrastructure implications (equipment, laboratory layout, safety installations)?	23
	What are the procurement implications (reagents, consumables, documentation)?	28
	What are the implications for quality assurance (internal and external)?	17
Layer 4 Scale-up analysis	What are the projected impacts of going to scale with the new tool? 1 Cost savings to patients in relation to income 2 Cost savings to health providers/ the health system 3 Effects on transmission of improved infection control as a result of the new tool	18
Layer 5 Policy analysis	What other similar technologies are available or likely to become available?	29
	How do similar existing or emerging technologies compare in their projected performance within each of the layers above?	25

Layer 1: Effectiveness analysis

This layer requires evidence about the accuracy (sensitivity and specificity) of new tools and approaches, but also flags the need to go further than this and build evidence on effectiveness. Data on sensitivity and specificity are universally provided by developers of new diagnostics, and their positive and negative predictive values have been suggested by GRADE as proxies for patient-important outcomes in the assessments of new tools. However, estimations of the number of patients who might start and complete appropriate treatment are typically calculated by extrapolating these parameters, rather than relying on evidence from field trials to provide estimates of actual numbers. All too often, diagnostic evaluations assess new tests solely in terms of their diagnostic potential (accuracy), which may not always translate into appropriate clinical or public health management decisions for patients within the context of health services (effectiveness).

Layer 2: Equity analysis

This layer examines who benefits from the new intervention. The Global Plan to Stop TB highlights the need to 'prioritise the needs of the poor and vulnerable', recognising that the greatest burden of TB is found among poor people, who also face the greatest barriers in access to care.[22] Typically, however, the systematic measurement of equity in health and health interventions is either absent or sporadic. Although the first MDG is expressed in terms of an equitable outcome, the health and other goals that are intended to contribute to this make no specific reference to equity or distributional issues.[30]

Layer 3: Health systems analysis

This layer examines the health systems requirements of a new intervention, for example human resources, infrastructure, operating procedures, quality assurance, procurement and maintenance.

These data are sometimes collected during the demonstration studies (Figure)—studies in optimised operational settings—of new diagnostics, but not in all cases. Even where they are collected, the improvement to operations necessarily provided through the demonstration study may mask issues that become apparent in implementation ('real world') studies. This layer provides crucial data for assessing the feasibility of implementation and for identifying where key constraints, or bottlenecks, in the system may occur.

Layer 4: Scale-up analysis

This layer projects and models the full economic costs as well as the clinical and epidemiological effects of going from demonstration or implementation studies to full scale (national or regional) with a new intervention. Health system, patient and societal

perspectives are all important here. Modelling techniques can provide information concerning the epidemiological benefits of scaling up and, when combined with patient costs from Layer 2, total additional costs or savings for patients. At the same time mathematical systems analysis techniques can outline the potential constraints to and resources required for scale-up. When combined with cost analyses from Layer 3, these can give an indication of total resources required as well as identify and quantify likely resource gaps.

Layer 5: Policy analysis

This layer critically appraises the new intervention studied in Layers 1–4 against other interventions that are available or may become available for uptake in the short to medium term. An important part of this layer is a scoping of the risk that a given new diagnostic test may be supplanted by newer technology within a short period of time. It requires a rapid assessment of data on other pipeline diagnostics from the previous four layers and a review of whether changes made for one diagnostic may provide a better platform for the next technology or, alternatively, whether the new technology is 'disruptive',[31] or 'market transformational',[32] both terms used to describe a technology that could radically alter the way in which TB diagnosis is achieved.

USING THE IAF

The IAF can be used by diagnostics research teams during the 'demonstration' and 'evidence for scale-up, delivery and access' phases of development shown in the Figure. The latter may take the form of field evaluation, or implementation, studies in non-optimised settings, or of other operational research activities. The framework can also be used by international policy makers during the policy development process to systematically assess a broader range of evidence, and by national policy makers to support adoption, implementation and scale-up decisions.

The IAF has already been used for the development of protocols for a multi-country research programme to study the implementation of line-probe assays (LPAs), which were recommended by WHO STAG-TB in 2008.[33] Representatives from three countries (Russia, Brazil and South Africa), all clinicians or laboratory specialists, with other members of the TREAT TB core group, discussed the priority research questions they would like to answer with regard to the use of LPAs, and mapped these questions to each layer of the IAF. All the questions raised mapped to one layer of the framework, and all layers were addressed; the resulting framework is shown in Table 2. Each of these teams now has a different protocol for collecting the evidence, due to the stage at which their NTPs are with regard to rolling out LPA. Nevertheless, each will provide data against the same set of outcome indicators, facilitating comparisons across different epidemiological settings.

Table 2 Use of the IAF for designing LPA field studies

Layer of assessment	Kinds of questions being addressed: questions and issues raised by multi-country research teams
Layer 1 Effectiveness analysis	How many additional cases will be identified who would otherwise not have been identified? How many additional cases will actually start treatment/achieve cure/avoid death as a result of using LPAs? What will be the effect on tuberculosis transmission? How will LPA affect the timeliness in results influencing a clinical or treatment decision?
Layer 2 Equity analysis	Who is benefiting from LPA implementation and why? Is the test sufficiently accurate for all patients? What are the risks to patients/others? What costs will patients face? How acceptable is the test to patients? Are there inequalities in access to LPA?
Layer 3 Health system analysis	What is the effectiveness and/or efficiency from a health system perspective? What effect will LPA have on how cases are managed in the health system? What quality assurance mechanisms need to be in place? What information systems need to be in place? What are the human resource requirements in the health system? What are the laboratory issues (including infrastructure, e.g., utilities, space; personnel, e.g., numbers and skills; monitoring system for laboratory)? How will the challenge of mixed infections be addressed? What are the safety issues? How will the results be interpreted and standardised?
Layer 4 Scale-up analysis	What are the obstacles to the rollout? What are the human resource and training requirements for full national scale-up?
Layer 5 Policy analysis	How does LPA compare with conventional 'old' methods vs. other new methods that may be available in the short to medium term? How does LPA interface with other existing and new diagnostics that will be recommended and implemented in the future (e.g., GeneXpert)?[34] Should routine drug susceptibility testing be completely dropped and replaced by LPA?

IAF = impact assessment framework; LPA = line-probe assay.

The central methodology that we advocate to feed robust evidence into Layers 1–3 is the prospective randomised controlled trial (RCT). This design permits comparison between the existing technology and approach (control) and the new (intervention), as follows:

For Layer 1, a comparison of effects on 1) numbers of patients achieving important outcomes (including diagnosis, start of treatment and treatment completion), and 2) time to achieving these outcomes.

For Layer 2, a comparison of effects on different patient sub-groups (e.g., poor vs. less poor, adults vs. children). Equity may be assessed based on outcome indicators among different groups, in terms of morbidity or mortality measures, or process indicators such

as health service use.[30] Analysis of socio-economic status may use asset-based measures to define different socio-economic groups.[35] Demographic and health surveys and more recent TB prevalence surveys are increasingly using these methods.[36,37]

For Layer 3, a comparison of the health system inputs is required. Data for this may be gained through economic analyses of standard vs. new diagnostic interventions, focusing on the health system and not just the tool, and through interviews with health systems personnel.

Data for these comparisons can be obtained across all study participants in both intervention and control arms, or through nested sub-studies on more limited numbers. For example, in-depth qualitative and quantitative research on patient costs incurred during a diagnostic process (either control or intervention) is time consuming, and data are thus only collected for a subgroup of study participants. Data from Layers 1–3 can then be fed into the modelling and other methodologies required in Layers 4 and 5.

We recognise that the type of randomised trial employed will depend on the stage of diagnostic development to which the IAF is being applied. During demonstration studies (which may be conducted prior to STAG-TB approval), an explanatory RCT with well-controlled study conditions and data collection instruments is appropriate. During subsequent implementation or operational research, a pragmatic RCT (PRCT) approach using existing health system data will be more suitable (for a fuller description of the difference between explanatory and pragmatic RCTs see Zwarenstein et al.[38]) There are concerns that RCT designs deny some patients (those in the control arm) the assumed benefits of a new technology —especially in the implementation research of STAG-approved technologies. Such ethical concerns need to be addressed, for example by ensuring that the PRCT includes a scale-up plan, such as through a step wedged approach in which all sites access the technology, but in a phased manner, to allow for comparisons between those with and those without the technology.

NEXT STEPS AND OVERCOMING CHALLENGES TO USING THE IMPACT ASSESSMENT FRAMEWORK

The framework will continue to be revised as experience in using it for research design and implementation continues. It will have value for other diagnostics tools and also for drugs and vaccines. The research methodologies for addressing each of the different layers are under constant development. As the multidisciplinary research teams needed to implement these methodologies are currently uncommon in many countries, capacity building involving training, mentoring and partnership between service delivery programmes, academic organisations and patient organisations will be required. The increased focus on patient-centred outcomes in particular will provide opportunities for patient representatives and organisations to become more engaged in the research process. If patient groups are empowered to collect and analyse relevant data—particularly in Layer 2—it will give them a greater voice in policy decision-making at the national and international levels.

When it was originally developed, the IAF was envisaged as being applicable to single new tests. However, as we move further into implementation research it is clear that it will also need to be applied to packages of tests, or combinations of existing and novel tests, along with all the additional inputs required to introduce such packages and combinations in different algorithms; this challenge is currently being addressed under the TREAT TB initiative.

The questions in the different layers of the IAF do not all necessarily carry equal weight in any given circumstance. For example, a new test for detecting drug resistance that is best suited for deployment in a central reference laboratory may be more important in monitoring drug resistance patterns than in directly improving patient access. The questions about which patient group or type of patient benefits (Layer 2) may then assume lesser importance, whereas these may be key research questions in a diagnostic approach or test that is aimed at 'point of care'.

We also recognise that while the IAF provides a body of evidence for policy makers, evidence alone is often not the only driver of policy change; process, context and sometimes subjective factors, for example expert or political opinion, can also play a substantial role. These factors need clearer and more systematic documentation and analysis in the process of implementation research. This is the subject of a forthcoming study by Bissell et al.

There are concerns that accumulating a comprehensive evidence base such as the one we advocate here will take too long and be too costly; rather than promoting the rational uptake of new technologies, it will instead impede the introduction of much needed innovations. Such concerns are valid, but they must be balanced against the dangers of the premature introduction of tools into unprepared and under-resourced health systems, often as a result of lobbying or forceful marketing. To counter both sides of this argument, research and implementation partners need to come together and collaborate on an unprecedented scale, and with a renewed sense of urgency. By directly addressing the concerns of policy makers through the research process, the adoption and implementation of new tools should be achieved more rapidly, sustainably and with beneficial effects for affected populations.

Acknowledgements

The authors thank the New Diagnostics Working Group (NDWG) for funding the initial development of this framework and the United States Agency for International Development for funding the TREAT TB (Technology, Research, Education and Technical Assistance for TB) initiative, which has enabled its evolution. They also thank all those people who have contributed to the process of developing the impact assessment framework outlined in this article; these include R Thomson, R Beddell, M Van Lettow, A Harries, R Dacombe, members of the NDWG and members of the TREAT TB management team.

References

1 Stop TB Partnership, World Health Organization. The global plan to stop TB 2006–2015. Geneva, Switzerland: WHO, 2006.
2 World Health Organization Stop TB Partnership New Diagnostics Working Group. Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics. Geneva, Switzerland: WHO, 2009. http://whqlibdoc.who.int/publications/2009/9789241598811_eng.pdf Accessed September 2010.
3 World Health Organization Special Programme for Research and Training in Tropical Diseases/Foundation for Innovative New Diagnostics. Diagnostics for tuberculosis: global demand and market potential. Geneva, Switzerland: WHO, 2006.
4 World Health Organization. Moving research findings into new WHO policies. Geneva, Switzerland: WHO, 2008. http://www.who.int/tb/dots/laboratory/policy/en/index4.html Accessed September 2010.
5 Guyatt G H, Oxman A D, Vist G, et al. Rating quality of evidence and strength of recommendations GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336: 924–926.
6 Guyatt G H, Oxman A D, Kunz R, et al. Rating quality of evidence and strength of recommendations: what is 'quality of evidence' and why is it important to clinicians? BMJ 2008; 336: 995–998.
7 Schunemann H J, Oxman A D, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336: 1106–1110.
8 Guyatt G H, Oxman A D, Kunz R, et al. Rating quality of evidence and strength of recommendations: incorporating considerations of resources use into grading recommendations. BMJ 2008; 336: 1170–1173.
9 Guyatt G H, Oxman A D, Kunz R, et al. Rating quality of evidence and strength of recommendations: going from evidence to recommendations. BMJ 2008; 336: 1049–1051.
10 Pai M, Ramsay A, O'Brien R. Evidence-based tuberculosis diagnosis. PLoS Med 2008; 5: 1043–1049.
11 Pai M, Minion J, Steingart K, Ramsay A. New and improved tuberculosis diagnostics: evidence, policy, practice, and impact. Curr Opin Pulm Med 2010; 16: 271–284.
12 Pai M, Ramsay A, O'Brien R J. Comprehensive new resource for evidence-based TB diagnosis. Expert Rev Mol Diagn 2009; 9: 637–639.
13 Marais B J, Raviglione M C, Donald P R, et al. Scale-up of services and research priorities for diagnosis, management, and control of tuberculosis: a call to action. Lancet 2010; 375: 2179–2191.
14 Stop TB Partnership, World Health Organization Retooling Task Force. New technologies for tuberculosis control: a framework for their adoption, introduction and implementation. WHO/HTM/STB/2007.40. Geneva, Switzerland: WHO, 2007.
15 Organisation for Economic Cooperation and Development–Development Assistance Committee (OECD-DAC). Glossary of key terms in evaluation and results-based management proposed harmonized terminology. 2002. Paris, France: OECD-DAC, 2002. http://www.undg.org/index.cfm?P=224 Accessed September 2010.
16 Wilkinson D, Newman W, Reid A, Squire S B, Sturm A W, Gilks C F. Trial-of-antibiotic algorithm for the diagnosis of tuberculosis in a district hospital in a developing country with high HIV prevalence. Int J Tuberc Lung Dis 2000; 4: 513–518.
17 Mundy C J F, Harries A D, Banerjee A, Salaniponi F M L, Gilks C F, Squire S B. Quality assessment of sputum transportation, smear preparation and AFB microscopy in a rural district in Malawi. Int J Tuberc Lung Dis 2002; 6: 47–54.
18 Mundy C J F, Bates I, Nkhoma W, et al. The operation, quality and costs of a district hospital laboratory service in Malawi. Trans R Soc Trop Med Hyg 2003; 97: 403–408.
19 Squire S B, Belaye A K, Kashoti A, et al. 'Lost' smear-positive pulmonary tuberculosis cases: where are they and why did we lose them? Int J Tuberc Lung Dis 2005; 9: 25–31.
20 Ramsay A, Squire S B, Siddiqi K, Cunningham J, Perkins M. The bleach microscopy method and case detection for TB control. Int J Tuberc Lung Dis 2006; 10: 256–258.
21 Cambanis A, Yassin M A, Ramsay A, Squire S B, Arbide I, Cuevas L E. A one-day method for the diagnosis of pulmonary tuberculosis in rural Ethiopia. Int J Tuberc Lung Dis 2006; 10: 230–232.
22 Kemp J R, Mann G H, Nhlema Simwaka B, Salaniponi F M L, Squire S B. Can Malawi's poor afford free TB services? Patient and household costs associated with a TB diagnosis in Lilongwe. Bull World Health Organ 2006; 85: 580–585.
23 Liu X, Thomson R, Gong Y, et al. How affordable are TB diagnosis and treatment in rural China? An analysis from community and TB patient perspectives. Trop Med Int Health 2007; 12: 1464–1471.
24 Theobald S, Taegtmeyer M, Squire S B, et al. Towards building equitable health systems in sub-Saharan Africa: what can we learn from case studies on operational research? Health Res Pol Syst 2009; 7: 26.
25 Ramsay A, Cuevas L E, Mundy C, et al. New policies, new technologies: modelling the potential for improved smear microscopy services in Malawi. PLoS ONE 2009; 4: e7760.
26 World Health Organization, The Stop TB Partnership. The Stop TB strategy: building on and enhancing DOTS to meet the TB-related Millennium Development Goals. Geneva, Switzerland: WHO, 2006.
27 Yagui M, Perales M T, Asencios L, et al. Timely diagnosis of MDR-TB under program conditions: is rapid drug susceptibility testing sufficient? Int J Tuberc Lung Dis 2006; 10: 838–843.
28 Ramsay A, Bonnet M, Gagnidze L, Githui W, Varaine F, Guerin P J. Sputum sex and scanty smears: new case definition may reduce sex disparities in smear-positive tuberculosis. Int J Tuberc Lung Dis 2009; 13: 613–619.
29 Harries A D, Michongwe J, Nyirenda T E, et al. Using a bus service for transporting sputum specimens to the Central Reference Laboratory: effect on the routine TB culture service in Malawi. Int J Tuberc Lung Dis 2004; 8: 204–210.
30 Gwatkin D R. Health inequalities and the health of the poor: what do we know? What can we do? Bull World Health Organ 2000; 78: 3–18.
31 Christensen C M. The innovator's dilemma: when new technologies cause great firms to fail. Boston, MA, USA: Harvard Business School Press, 1997.
32 Blumstein C, Goldstone S, Lutzenhiser L. A theory-based approach to market transformation. Energy Policy 2000; 28: 137–144.
33 World Health Organization. New WHO policies. Geneva, Switzerland: WHO, 2010. http://www.who.int/tb/dots/laboratory/policy/en/index.html Accessed September 2010.
34 Blakemore R, Story E, Helb D, et al. Evaluation of the analytical performance of the Xpert MTB/RIF Assay. J Clin Microbiol 2010; 48: 2495–2501.
35 Wagstaff A. Socio-economic inequalities in child mortality:

comparisons across developing countries. Bull World Health Organ 2000; 78: 19–29.
36 Lönnroth K, Holtz T H, Cobelens F, et al. Inclusion of information on risk factors, socio-economic status and health seeking in a tuberculosis prevalence survey. Int J Tuberc Lung Dis 2009; 13: 171–176.
37 Muniyandi M, Rajeswari R. Socio-economic inequalities of tuberculosis in India. Expert Opin Pharmacother 2008; 9: 1623–1628.
38 Zwarenstein M, Treweek S, Gagnier J J, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008; 337: a2390.

RÉSUMÉ

La nécessité d'une base de preuves solide et complète pour servir à la prise de décisions en ce qui concerne la mise en œuvre d'outils de diagnostic et d'approches nouvelles et améliorées a été soulignée par un certain nombre de responsables ; parmi ceux-ci, des membres du groupe de travail sur les nouveaux outils diagnostiques New Diagnostics Working Group (NDWG) et du sous-groupe pour l'introduction d'approches et d'outils nouveaux (Subgroup for Introducing New Approaches and Tools) du Partenariat Stop TB. Afin de rassembler ces évidences de manière systématique, nous avons élaboré un réseau d'évaluation d'impacts (IAF) qui fait le lien entre apports et résultats finaux.

L'IAF comporte cinq couches interconnectées : analyse d'efficience, analyse d'équité, analyse des systèmes de santé, analyse de l'extension et analyse de la politique. Il peut être utilisé par ceux qui élaborent de nouvelles techniques de diagnostic et par d'autres équipes de recherche intéressées à rassembler autant de données possibles en rapport avec la politique à suivre avant, pendant et après la phase de démonstration de l'élaboration de l'outil. Les évidences rassemblées peuvent être utilisées par les décideurs politiques internationaux et nationaux pour soutenir des décisions d'adoption, de mise en œuvre et d'extension. L'initiative TREAT TB (Technologie, Recherche, Education et Assistance Technique) utilise l'IAF dans sa recherche opérationnelle et dans ses évaluations sur terrain des nouveaux outils et des nouvelles approches du diagnostic de la tuberculose ; l'IAF a été incorporé dans la publication récente du NDWG : « Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics ». Cet article décrit l'IAF et les processus employés pour son amélioration et suggère les étapes ultérieures pour surmonter les défis que comporte sa mise en œuvre.

RESUMEN

Varios interesados directos, entre ellos el Grupo de Trabajo sobre Nuevos Diagnósticos (NDWG) y el subgrupo de introducción de nuevos enfoques e instrumentos de la Alianza Alto a la Tuberculosis, han destacado la necesidad de contar con una base de datos científicos sólida y exhaustiva, a fin de respaldar la toma de decisiones relacionadas con la introducción de nuevos enfoques e instrumentos perfeccionados de diagnóstico. Con el objeto de recoger estos datos de manera sistemática, se ha diseñado un marco de evaluación del impacto (IAF), que vincula los datos aportados con los resultados obtenidos.

El marco de evaluación del impacto comporta cinco estratos interconectados: el análisis de eficacia, el análisis de equidad, el análisis de los sistemas de salud, el análisis de la ampliación de escala y el análisis de las políticas. Este marco pueden usarlo los creadores de nuevos métodos diagnósticos y otros grupos científicos interesados, durante el desarrollo de un nuevo instrumento con el fin de recoger la máxima cantidad de datos pertinentes a las políticas, antes de la fase de demostración, durante la misma o después de ella. Los datos científicos recogidos pueden ser útiles a los encargados de definir las políticas a escala nacional o internacional, a fin de respaldar las decisiones de adopción, ejecución o ampliación de escala. En la iniciativa TREAT TB (Tecnología, Investigación, Educación y Asistencia Técnica para la Tuberculosis) se aplica el IAF en las evaluaciones de la investigación operativa y de terreno de los nuevos instrumentos y estrategias utilizados en el diagnóstico de la tuberculosis. También se ha incorporado en la publicación reciente del NDWG: 'Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics' (Estrategias encaminadas a mejorar los métodos diagnósticos de la tuberculosis: un plan de acción para el desarrollo de medios diagnósticos). En el presente artículo se describe el IAF y los mecanismos que permiten mejorarlo y se sugieren nuevos pasos que contribuyan a superar las dificultades que plantea su ejecución.

Tuberculosis 4

Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice

Robert S Wallis, Madhukar Pai*, Dick Menzies, T Mark Doherty, Gerhard Walzl, Mark D Perkins†, Alimuddin Zumla†*

Lancet 2010; 375: 1920–37
Published Online
May 19, 2010
DOI:10.1016/S0140-6736(10)60359-5

See **Comment** page 1852

This is the fourth in a **Series** of eight papers about tuberculosis

*These authors contributed equally and are joint first authors

†Joint senior authors

Pfizer, New London, CT, USA (R S Wallis MD); **Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada** (M Pai MD, D Menzies MD); **Montreal Chest Institute, Montreal, QC, Canada** (M Pai, D Menzies); **Statens Serum Institute, Copenhagen, Denmark** (Prof T M Doherty PhD); **Stellenbosch University, Stellenbosch, South Africa** (Prof G Walzl PhD); **Foundation for Innovative New Diagnostics, Geneva, Switzerland** (M D Perkins MD); **and Department of Infection, University College London Medical School, London, UK** (Prof A Zumla FRCP)

Correspondence to:
Prof Alimuddin Zumla, Centre for Infectious Diseases and International Health, Department of Infection, University College London Medical School, Windeyer Institute of Medical Sciences, 46 Cleveland Street, London W1T 4JF, UK
a.zumla@ucl.ac.uk

or Dr Mark D Perkins, Foundation for Innovative New Diagnostics, Avenue de Budé 16, 1202 Geneva, Switzerland
mark.perkins@finddiagnostics.org

For more on **Evidence-based TB Diagnosis** see http://www.tbevidence.org

Human infection with *Mycobacterium tuberculosis* can progress to active disease, be contained as latent infection, or be eradicated by the host response. Tuberculosis diagnostics classify a patient into one of these categories. These are not fixed distinct states, but rather are continua along which patients can move, and are affected by HIV infection, immunosuppressive therapies, antituberculosis treatments, and other poorly understood factors. Tuberculosis biomarkers—host or pathogen-specific—provide prognostic information, either for individual patients or study cohorts, about these outcomes. Tuberculosis case detection remains difficult, partly because of inaccurate diagnostic methods. Investments have yielded some progress in development of new diagnostics, although the existing pipeline is limited for tests for sputum-smear-negative cases, childhood tuberculosis, and accurate prediction of reactivation of latent tuberculosis. Despite new, sensitive, automated molecular platforms for detection of tuberculosis and drug resistance, a simple, inexpensive point-of-care test is still not available. The effect of any new tests will depend on the method and extent of their introduction, the strength of the laboratories, and the degree to which access to appropriate therapy follows access to diagnosis. Translation of scientific progress in biomarkers and diagnostics into clinical and public health programmes is possible—with political commitment, increased funding, and engagement of all stakeholders.

Introduction

Tuberculosis, although a curable disease, continues to be one of the most important infectious causes of death worldwide. Despite substantial investments and progress made in expansion of the directly observed therapy, short course (DOTS) strategy and improved treatment completion rates, inadequate case detection remains a major obstacle to global control of tuberculosis. Efforts during the past decade to consistently diagnose and treat the most infectious cases have slowed the rate of disease incidence, but have not yielded substantial progress towards elimination. This experience has refocused attention on research and development for improved diagnostics, therapeutics, and vaccines—areas in which progress has historically been slow. Human *Mycobacterium tuberculosis* infection is almost always acquired by inhalation of infected aerosol droplets, which are generated by people with active pulmonary disease

Search strategy and selection criteria

For the tuberculosis biomarker section, we searched publications in PubMed and Google Scholar (1995–2009), the Cochrane library (2001–09), and Embase (2001–06) with the terms "tuberculosis", "*Mycobacterium tuberculosis*", "biomarkers", "diagnostics", and "clinical trials". For the section on tuberculosis diagnostics, the search strategy was a 10-year review of diagnostic studies in PubMed and Embase. This search was supplemented by searching the website Evidence-based TB Diagnosis by the Stop TB Partnership's New Diagnostics Working Group. We searched with the terms "tuberculosis", "*Mycobacterium tuberculosis*", "diagnosis", "diagnostics tests", and "accuracy". For both sections, we mainly selected publications in the past 10 years, but did not exclude commonly referenced and highly regarded older publications. We also reviewed studies cited by articles identified by this search strategy, and selected those that we regarded as relevant. Review articles are preferentially cited to provide readers with more details and references than this overview can accommodate.

Key messages

- Diagnostics classify patients at one point in time, whereas biomarkers can provide prognostic information about future health status and can advance knowledge of disease pathogenesis.
- Qualified tuberculosis biomarkers are most urgently needed as predictors of reactivation and cure, and indicators of vaccine-induced protection. The biomarker most closely approaching qualification is 2-month culture conversion as a predictor of relapse risk.
- The tuberculosis diagnostics pipeline has rapidly grown, with development of several promising technologies.
- The existing tuberculosis diagnostics pipeline still does not have a simple, rapid, inexpensive point-of-care test. Accurate, rapid tests are also needed for smear-negative and childhood tuberculosis, as are tests for latent tuberculosis with increased predictive value for reactivation.
- Several diagnostics and diagnostic strategies have been endorsed by WHO and are being introduced into clinical use and national tuberculosis control programmes.
- Governments in all countries, especially industrialised countries, have to increase funding for tuberculosis research and control.

coughing (figure 1). However, the infection infrequently progresses directly to active disease, and is more often contained—at least initially—by the host immune response. The resulting latent infection can be eradicated, or can persist and reactivate many years later. Tuberculosis chemotherapy can also contain the disease, but leave a latent infection that is capable of causing relapse. This dynamic process can be started anew at any time by exogenous reinfection.

Tuberculosis diagnostics form the basis of classification of patients in this system. As diagnostic accuracy has increased, it has become apparent that these are not entirely distinct states, but instead represent gradations along which patients might move. Even within individual patients, foci of latent infection can coexist alongside sites of active *M tuberculosis* replication. Medical treatment, vaccine and immune status, and concomitant illness all affect this balance between host and pathogen, favouring one or another clinical outcome and thus representing the interface between prognostics and diagnostics. In this overview, we describe the development of tuberculosis biomarkers and diagnostics, knowledge gaps and scientific obstacles, and limitations of the existing pipeline of biomarkers and diagnostics, and summarise the major challenges in translation of scientific progress into action.

Biomarkers for tuberculosis

Biomarkers provide prognostic information about future health status, either for individual patients or cohorts in clinical trials. Biomarkers can thus indicate normal or pathogenic processes, or pharmacological responses to therapeutic intervention.[1] In clinical trials, biomarkers can form the basis of surrogate endpoints, which can substitute for a clinical endpoint based on epidemiological, therapeutic, pathophysiological, or other scientific evidence, thereby assisting candidate selection during drug discovery, accelerating dose selection in early clinical research, and shortening the time to licensing of new drugs and vaccines. In routine clinical care, biomarkers can allow stratification of individual patients according to outcome risks, thus easing targeted interventions that might not otherwise produce overall benefit. Biomarkers can also help to advance basic knowledge of disease pathogenesis.

The need for biomarkers in tuberculosis is most crucial in three areas: in patients with active disease, to predict durable (non-relapsing) treatment success; in patients with latent *M tuberculosis* infection, to indicate reactivation risk and predict treatment success; and in people other than those with active disease, to indicate protection from tuberculosis by new vaccines (panel 1).

Biomarkers predicting durable cure

The marker with which there is greatest experience as a predictor of non-relapsing cure is sputum culture status after 2 months of therapy. Wallis and colleagues[4] used meta-regression to examine these parameters in 30 paired study groups of 5500 patients in four regions worldwide. The analysis found that an incremental effect of a new treatment on relapse is highly likely to be captured as a corresponding change in culture conversion (figure 2; $p<0·0001$). This finding supports a role for 2-month sputum culture conversion in the accelerated approval of new tuberculosis drugs, potentially shortening the time needed for licensing of new drugs for multidrug-resistant (MDR) tuberculosis by many years. No other tuberculosis biomarker approaches this level of qualification. However, despite this compelling performance as a surrogate endpoint in clinical trials, sputum culture conversion is a poor prognostic marker for individual patients. One study noted, for example, that although 2-month culture positivity was an independent predictor of relapse for individuals (hazard ratio 2·8, 95% CI 1·7–4·7), its positive predictive value (18%) and sensitivity (50%) were low.[3] This apparent discordance between trial surrogacy and patient prognostication could arise from the practice of collecting sputum cultures only once per month, thus obscuring within-patient variability. Some relapses could also arise from bacterial subpopulations that are not readily detected in sputum by culture on solid medium.

Efforts to improve on these characteristics convert the binary endpoint of culture conversion to a continuous variable by measuring the rate of decline of viable *M tuberculosis* in sputum at several timepoints during the first 1–2 months of therapy, either as colony counts on agar or time to positivity in liquid culture.[5–8] One small trial[9] using serial counts identified moxifloxacin and gatifloxacin as superior to ofloxacin and ethambutol despite similar rates of 2-month culture conversion. However, three of four adequately powered trials of moxifloxacin did not show an effect on 2-month status, including the regimen that was indicated in mice as most likely to accelerate sterilisation.[102–105] A very large clinical trial (Rapid Evaluation of Moxifloxacin in the Treatment of Sputum Smear Positive Tuberculosis [REMOX-TB]) in

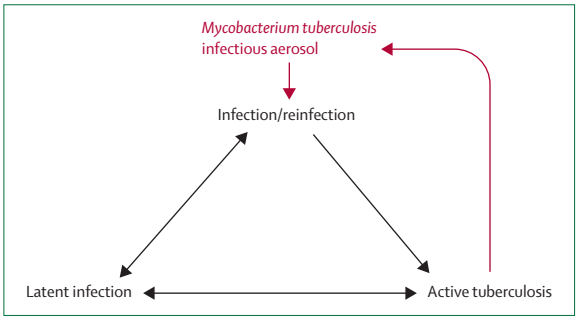


Figure 1: **Clinical stages or states of *Mycobacterium tuberculosis* infection**
Bidirectional movement between states can occur as a result of exogenous or endogenous effects, including inhalation of infected aerosol droplets, vaccination, antituberculosis chemotherapy, or concomitant illness such as HIV.

Panel 1: Candidate *Mycobacterium tuberculosis* and host tuberculosis biomarkers

Predication of durable (non-relapsing) tuberculosis cure

Microbial markers in sputum

- 2-month culture conversion[2–4]
- Serial colony counts or time to culture positivity[5–9]
- Other microbial markers[8,10–14]

Other microbial markers

- Urine *M tuberculosis* DNA,[15,16] lipoarabinomannan[17–26]
- Volatile organic compounds[27,28]

Mycobactericidal activity

- Whole blood culture[29–31]

Tuberculosis-specific T-cell function

- Interferon γ,[32–37] interleukin 4δ2 splice variant[38–40]

Macrophage activation markers

- Neopterin,[41–45] procalcitonin,[46–53] C-reactive protein,[54–59] soluble intercellular adhesion molecule 1,[60–64] soluble urokinase plasminogen activator receptor,[64,65] monocyte CDllc[66]

Multiple host markers

- Proteomics[67,68]
- Transcriptomics[69,70]

Indication of reactivation risk and prediction of eradication of latent infection

Tuberculosis-specific T-cell function

- Interferon γ[71–77]
- Interferon-induced protein 10[78–80]
- Interleukin 4δ2 splice variant[38–40,81,82]
- Skin test[83,84]

Macrophage activation

- Neopterin[85]
- Procalcitonin[51]

Prediction of vaccine efficacy

Tuberculosis-specific T-cell function

- Interferon γ[86]
- Polyfunctional T cells[87–90]

Mycobactericidal activity

- Whole blood culture[91–99]
- Mononuclear cells[86,98,100,101]

progress with relapse as its primary endpoint will probably help to resolve these contradictory findings.

Small studies have examined levels of *M tuberculosis* antigen 85 and 85B RNA in sputum during treatment. In one study,[8] the magnitude and duration of increases in this protein during the first week of treatment predicted relapse or failure in four of 42 patients. A second study[12] noted that 85B RNA was cleared more rapidly from sputum during therapy than viable colony counts, but did not predict subsequent relapse in one patient. Other factors associated with mycobacterial dormancy (including proteins, RNA species, or lipids) could have greater predictive value.[13,14]

An important shortcoming of all sputum biomarkers is their limited role in paucibacillary and paediatric tuberculosis, and their lack of usefulness in latent *M tuberculosis* infection. One study has reported the presence of small fragments of *M tuberculosis* IS6110 DNA in urine of 34 of 43 patients with tuberculosis but not in healthy controls.[15] None of the patients had overt renal tuberculosis. The DNA fragments, termed transrenal DNA (tr-DNA), are thought to arise because of apoptosis of host cells. The investigators have reported that none of the 20 patients who were positive at diagnosis remained positive after 2 months of therapy.[16] Other studies of urinary mycobacterial DNA using different methods have generally shown lower sensitivity.[106–109] None has examined paediatric samples. A urinary test that could serve as both diagnostic and prognostic would be an important advance in paediatric tuberculosis. Other studies have examined lipoarabinomannan and other mycobacterial markers in urine, also with varying degrees of sensitivity.[17–26] Detection of volatile organic compounds in the breath of patients with pulmonary tuberculosis has been reported.[28] No studies have reported the changes in these markers during treatment or established a relation to clinical outcome or to another surrogate endpoint. Further study of non-sputum microbial markers is an area of priority for tuberculosis research.

Measurement of bactericidal activity in whole blood culture after oral dosing of new tuberculosis drugs can help with selection of dose and dosing interval, and can identify compounds which, owing to their mechanism of action and pharmacokinetic profiles, can show additive or synergistic effects when combined. Such effects might not be predicted well from animal models because of differences in absorption and metabolism.[110,111] Whole blood bactericidal activity during tuberculosis treatment correlates with the rate of decline in sputum colony counts, is superior in patients whose sputum cultures convert by the second month of treatment, and is superior during the intensive (four-drug) phase of treatment.[31] Two studies[29,30] reported that regimens for drug-sensitive tuberculosis were better than were those for MDR tuberculosis. These findings suggest that the whole blood model could also help in the identification of efficacious multidrug regimens.

Macrophages are activated by *M tuberculosis* via interactions with toll-like receptors. Several blood markers of this activation might have roles as tuberculosis biomarkers. Neopterin, for example, is increased at diagnosis of disease in proportion to extent of disease; it decreases during and after treatment.[41–45] In a small sample of HIV-uninfected patients matched for extent of disease at baseline, increased neopterin concentrations after completion of treatment were associated with relapse.[44] Several other markers are also increased at baseline in proportion to disease extent and to decline with treatment, including soluble intercellular adhesion molecule (sICAM) 1,[60–63] C-reactive protein,[55,57–59] soluble urokinase plasminogen

activator receptor,[65] and procalcitonin.[46–53] In one study, a mathematical model including change in sICAM 1 during the first week of therapy predicted 2-month sputum culture conversion.[64] As a group, these assays are simple, inexpensive, widely available, and can be done on frozen plasma samples; as a result, they can be readily incorporated into clinical trials or treatment protocols. They seem to have greatest prognostic value when measured at or near the completion of therapy. Further research is needed to establish the sensitivity of these tests to predict tuberculosis reactivation or relapse, and the extent to which their lack of specificity for *M tuberculosis* infection confounds their interpretation.

Multiple biomarkers, when combined, can do substantially better than can any one marker. For example, a panel consisting of leptin, prolactin, osteopontin, insulin-like growth factor II, macrophage inhibitory factor, and CA-125 yielded 95·3% sensitivity and 99·4% specificity for diagnosis of ovarian cancer, for which measurement of CA-125 alone detects only 30–40% of early cases.[112] Increased specificity and high predictive value may similarly be achieved for otherwise non-specific tuberculosis biomarkers by measuring multiple parameters by proteomics, transcriptomics, and metabolomics.[10,68] Tuberculosis can be differentiated from other infectious and inflammatory diseases on the basis of proteomic fingerprinting study of serum by surface-enhanced laser desorption/ionisation time-of-flight (SELDI-ToF) mass spectrometry.[67] Serum amyloid A and transthyretin were among the candidate biomarkers identified. This analytical method can detect a very large number of peptides, although it is fairly insensitive. A related approach reduces the potential number of candidate molecules to a small set of small molecules termed the metabolome, representing metabolic intermediates, hormones, other signalling molecules, and secondary metabolites. Its main disadvantage is that several analytical methods are necessary to complete their characterisation.

Two reports suggest the feasibility of distinguishing various stages of *M tuberculosis* infection by gene expression microarray. One study[70] recorded many candidate genes that were differentially expressed by mononuclear cells of patients with tuberculosis, people with latent infection, and uninfected people. However, a small subset of these genes—lactoferrin, CD64, and the Ras-associated GTPase 33A—was sufficient for classification of the three groups. A second report identified signature profiles of nine genes in blood that could distinguish four groups: patients with active tuberculosis, those with latent infection, cured patients, and cured patients with several previous episodes of tuberculosis.[69] Similar studies undertaken during or at completion of therapy could identify profiles associated with durable cure or relapse. However, the genomics and proteomics platforms might be susceptible to biases indicating regional differences in host and microbial genetics. These findings have yet to be verified across several clinical populations.

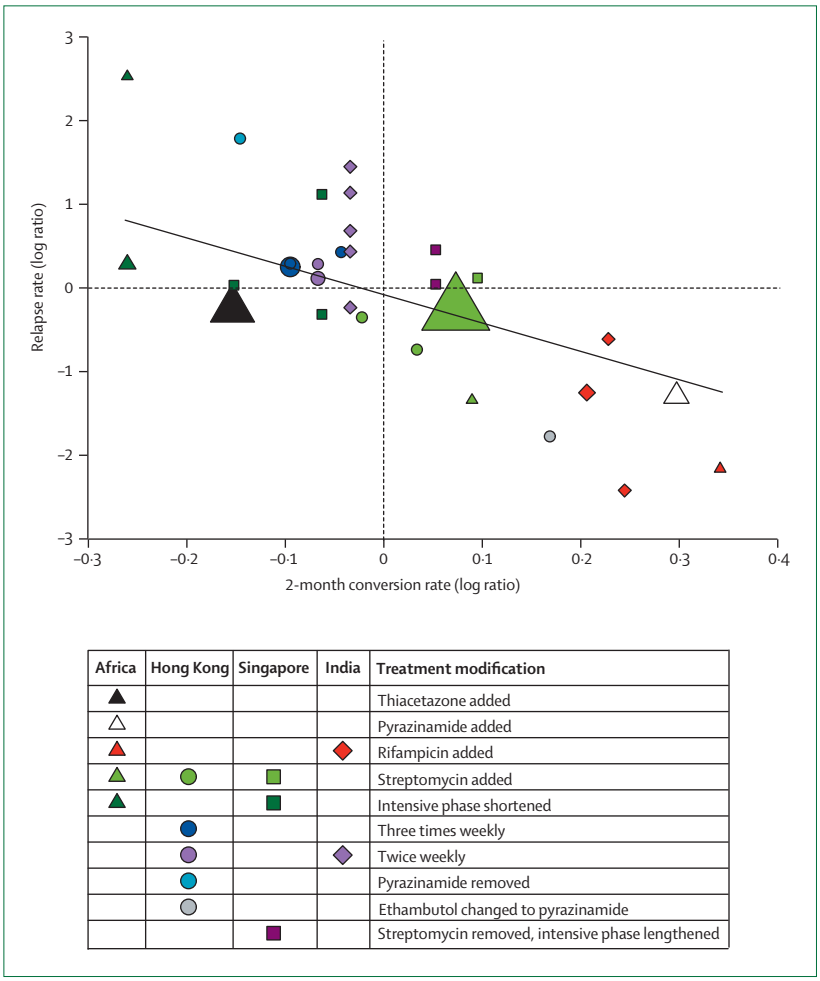


***Figure 2*: Relation between effects of new tuberculosis treatments on rates of relapse and 2-month sputum culture conversion in randomised controlled trials**
Axes indicate natural log rate ratios (experimental/control), with dotted lines indicating equality (no effect). Symbols indicate pairs of study groups, differing only in the intensive phase (n=16) or throughout treatment (n=14). Symbol sizes vary according to precision. The solid line indicates the results of meta-regression analysis (p<0·0001). Reproduced from reference 4, with permission from author and publisher.

T-cell-based assays of interferon-γ release for diagnosis of latent tuberculosis infection tend to show high levels in people with active disease at diagnosis that decrease after completion of treatment, but this pattern does not occur consistently.[32] A small study using a non-commercial assay to monitor responses at earlier timepoints noted that of 18 active cases of tuberculosis with positive T-cell responses at baseline, only five who did not show a microbiological or clinical response remained positive after 3 months of treatment.[33] Subsequent studies using commercial assays have not yielded such definitive results. Although most studies have concurred in finding sustained positive T-cell responses in tuberculosis treatment failures, most have found reversion of T-cell responses in responders to be too incomplete and delayed to be useful as a biomarker.[34–37,113] The diagnosis of active tuberculosis can also be established on the basis of T-cell frequencies at the site of disease rather than in blood.[114–116] However, the requirement for an

invasive procedure restricts the feasibility of this approach for diagnosis and treatment monitoring. Lastly, antibody concentrations to some mycobacterial antigens are raised at diagnosis and might be modulated by treatment; however, performance characteristics seem inadequate for a prognostic role.[117,118] Immunological memory seems to hamper the ability to quickly detect treatment effects, as is the case with T-cell assays.

Biomarkers indicating reactivation risk

Several natural history studies of household contacts of active cases of tuberculosis suggest that in HIV-uninfected people, particularly high or increasing concentrations of tuberculosis-specific interferon-γ production might predict overt tuberculosis, although the numbers of tuberculosis cases in these studies are small.[75–77] Positive responses to interferon-γ release assays (IGRAs) otherwise seem to confer only a small risk of reactivation (10–20 per 1000 person-years), which is similar to that of a positive tuberculin skin test. Some studies have suggested that relative mRNA concentrations of interferon γ, interleukin 4, and interleukin 4δ2 (a splice variant of interleukin 4) might be better predictors than interferon γ alone, since ratios of interferon γ or interleukins 4 and 4δ2 fall as healthy contacts develop tuberculosis, and increase as patients with tuberculosis are cured.[38–40,81] The ratio of interleukin 4 to 4δ2 is also increased in longstanding latent tuberculosis infection, presumably suggesting low risk of reactivation.[82] One study[85] reported finding intermediate concentrations of neopterin in health-care workers who were heavily exposed to tuberculosis, potentially indicating risk of reactivation of latent *M tuberculosis* infection. No studies have yet examined macrophage and T-cell markers together in this context.

Findings from studies in experimentally infected guineapigs suggest that prognostication of tuberculosis with the tuberculosis-specific antigens ESAT 6 and CFP 10 as skin tests might also be possible.[83] A similar though less pronounced occurrence had been described in patients with responses to a tuberculin skin test.[84] Such a skin test might show better specificity for latent *M tuberculosis* infection and increased positive predictive value for tuberculosis than might the tuberculin skin test. Confirmation of these findings in human beings and validating their prognostic significance are priority areas of research.

How the loss of CD4 T cells due to HIV infection will affect prognostication of tuberculosis with use of T-cell-based assays is unclear. One study of acute HIV infection noted that tuberculosis-specific T-cell interferon-γ responses were lost rapidly in four of five patients, all of whom remained well.[119] In the fifth, tuberculosis-specific responses increased progressively after HIV infection was acquired, culminating in the diagnosis of active tuberculosis. Increasing counts might correlate with antigen burden and presage tuberculosis reactivation in HIV-positive and HIV-negative people, but studies in HIV-infected patients with specific ranges of CD4 T cells will be needed to confirm this observation.

Several studies of people recently exposed to tuberculosis showed that T-cell frequencies decrease after completion of isoniazid preventive therapy, although they infrequently reverted to negative.[120–123] However, in other studies,[124–126] responses were unaffected. Factors affecting the likelihood of IGRA reversion due to isoniazid preventive therapy could include the duration of infection, the type of assay, the magnitude of the response, or the risk of reinfection.[127] No studies have specifically examined the prognostic significance of reversion. IGRAs are unlikely to be adequate indicators of successful isoniazid preventive therapy. Multiplex assays assessing both T-cell and macrophage factors could prove useful.

Biomarkers predicting vaccine efficacy

There are no qualified biomarkers to indicate protection by new vaccines against tuberculosis. Although both natural infection and vaccination with *Mycobacterium bovis* BCG result in the acquisition of delayed-type hypersensitivity and expansion of antigen-specific interferon-γ-producing T-cell populations, the link between these responses and protection from disease is weak. In the case of BCG, for example, interferon-γ-producing T-cell frequencies poorly predict the protective efficacy of various BCG strains in animals.[128] Protection might instead correlate better with the presence of polyfunctional antigen-specific T cells that secrete several cytokines, as has been described in leishmaniasis.[129] However, data for the use of this biomarker for vaccine-induced protection in tuberculosis are scarce. The potential effect of this insufficient knowledge is shown in two studies of the effect of route of BCG administration on its efficacy. The first study[89] noted superior immunogenicity in infants (interferon-γ, tumour necrosis factor, and interleukin-2 responses in both CD4 and CD8 T cells) when the vaccine was administered percutaneously rather than intradermally. However, a subsequent study[130] in this population showed the two methods of administration did not differ in protection against tuberculosis. These findings suggest that assessment of T-cell responses alone might be insufficient to predict protection from tuberculosis by vaccination.

For all other licensed vaccines, bactericidal or viral neutralisation assays have supplemented standard measurements of immunogenicity during development. Bactericidal assays have been described for *M tuberculosis* with mononuclear cell or whole blood culture.[86,91,92,101] Immune control of growth in these assays is inferior in people with negative tuberculin skin test and in young children; improved by BCG vaccination or vitamin D; impaired by chemokine receptor blockade, T-cell depletion, or HIV infection; restored by antiretroviral treatment; and might be strain specific.[86,91–99,131] Their predictive role for new tuberculosis vaccines has yet to be studied.

Effect of biomarkers on development timelines

The potential effect of biomarkers on the time and costs of development of new tuberculosis drugs and vaccines can be substantial. In the USA, Federal regulations (subpart H of 21CFR314) allow accelerated approval of new drugs for serious or life-threatening illnesses on the basis of a surrogate endpoint that is "reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence, to predict clinical benefit".[132] In 2009, an Advisory Committee convened by the US Food and Drug Administration (FDA) recommended nearly unanimously in favour of accelerated approval of new drugs for MDR tuberculosis on the basis of sputum culture conversion. Such approval will shorten the time needed for licensing of new, more effective treatments to patients with MDR tuberculosis by as much as 3 years. This strategy might also be used in development of new regimens, rather than single compounds. Here, measurement of serial sputum colony-forming unit counts and whole blood bactericidal activity in trials of 1–4 weeks' duration can provide a seamless progression from preclinical studies through trials resulting in licensing with culture conversion. Such a development plan might reduce by as much as a decade the time required to have new regimens for MDR tuberculosis without cross-resistance to any existing tuberculosis drug.

Strategies for biomarker qualification

Although biomarkers have historically been widely used in drug development and medical practice, only recently have pathways been created to include them in the regulatory review process. In the USA, the impetus for this change came from the National Institutes of Health Road Map and the FDA Critical Path Initiative, both of which sought to introduce greater efficiency in drug research and development.[133,134] In this context, the term validation refers to assay performance characteristics (eg, how accurately urinary albumin is measured), whereas qualification refers to linkage to biological processes (eg, to what extent do increases in urine albumin predict aminoglycoside nephrotoxicity).[135] Biomarker review at the FDA includes a voluntary data submission that is examined by a biomarker qualification review team. Three categories of certainty are described: biological plausibility; prognosis of clinical outcomes in disease; and capturing differences in efficacy in clinical trials.[136] The first category could be described as appropriate only for exploratory purposes, whereas the third might be needed for registration of a new therapy or vaccine. Of all the markers described in this review, only 2-month sputum culture conversion falls into the third category. Reaching this level of certainty is particularly challenging in the case of new vaccines for tuberculosis, since there is no effective modern vaccine to which BCG might be compared for the purposes of biomarker qualification, and since innate genetic factors not amenable to modulation by vaccination might account for and be detected by biomarkers that predict risk of disease in natural history studies.

Tuberculosis diagnostics

Progress towards elimination of tuberculosis has remained elusive despite intensified standard measures of control. After a period of global acceleration in 2001–05, the case detection rate worldwide decelerated in 2006 and 2007, reaching 63% in 2007.[137] Thus, the target of a case detection rate of at least 70% by 2005 has not yet been achieved, and is unlikely to be met until 2014.[137]

Insufficient access to advanced diagnostic tests has contributed to this suboptimum performance. Even in 2010, national tuberculosis programmes in disease-endemic countries continue to rely largely on antiquated and inaccurate methods such as direct smear microscopy, solid culture, chest radiography, and tuberculin skin testing. There is no rapid, point-of-care test that allows early detection of active tuberculosis at health clinics. Diagnosis of smear-negative tuberculosis in adults infected with HIV and in children continues to pose substantial clinical challenges. Even existing diagnostics are not used to their full potential because of poor access to health care and failures in health-care delivery systems, including poor synergy between national HIV/AIDS and tuberculosis programmes. Diagnostic delays, misdiagnosis, and inadequate implementation of existing tests result in increased morbidity and mortality in patients, and allow continued transmission of tuberculosis.[138] These restrictions of present case detection approaches are starkly visible in countries with a high prevalence of HIV infection or MDR tuberculosis, or both.[139–141]

Barriers to development of new tuberculosis diagnostics

Market failure has been an important factor hindering the development of new diagnostics for tuberculosis. Industry tends to avoid developing and marketing products that will be mainly used for poor patients in resource-limited countries because such products will not generate profits.[142,143] When products are available, neither pricing nor performance is adapted for developing countries, and their potential benefits are effectively unavailable for patients and health-care providers who need them most.

Furthermore, health systems in developing countries are generally weak, making them unable to take advantage of tuberculosis diagnostics to achieve best possible performance, and to introduce new advances in diagnostic technologies. This situation is the result of poor management, insufficient financial resources, inadequate human resources, and poor laboratory capacity.[144] For example, rapid tests for malaria are a model of the type of assay widely needed for tuberculosis, but only a small proportion of patients receiving malaria treatment are tested.[145] Rapid tests for HIV infection are highly accurate, but undiagnosed HIV infection is very common, and a large proportion of HIV-infected individuals do not

present for HIV testing until late in infection.[146] Only about 10–20% of people infected with HIV in Africa are aware of their status.[144] Furthermore, less than 3% of people with HIV infection are screened for tuberculosis, and globally, only about 20% of notified tuberculosis patients are aware of their HIV status.[147] These estimates suggest that even existing diagnostic strategies are poorly implemented in many settings.

Knowledge gaps and scientific obstacles impeding progress

Our understanding of the biology of *M tuberculosis* and interactions with the human host is incomplete, and these knowledge gaps impede the development of biomarkers that can distinguish between latent and active tuberculosis, and distinguish active tuberculosis from other diseases, especially in HIV-infected adults and children.[148–150] Present tests for latent *M tuberculosis* infection do not adequately distinguish resolved from persistent infection, and are unable to efficiently identify individuals who are at highest risk of reactivation.[151–155] Studies into predictive value of IGRAs show only modest predictive ability, and several studies show similar (and rather low) rates of progression in people with positive tuberculin skin test and IGRA results.[156–160]

Other important knowledge gaps pertain to diagnosis of smear-negative tuberculosis in children and HIV-infected individuals, and rapid and accurate identification of resistance to second-line antituberculosis drugs. Although molecular markers have been identified and successfully used as rapid and accurate tests for isoniazid and rifampicin resistance, testing for the resistance that characterises extensively drug-resistant tuberculosis is on a less robust scientific footing than is testing for MDR tuberculosis.[161]

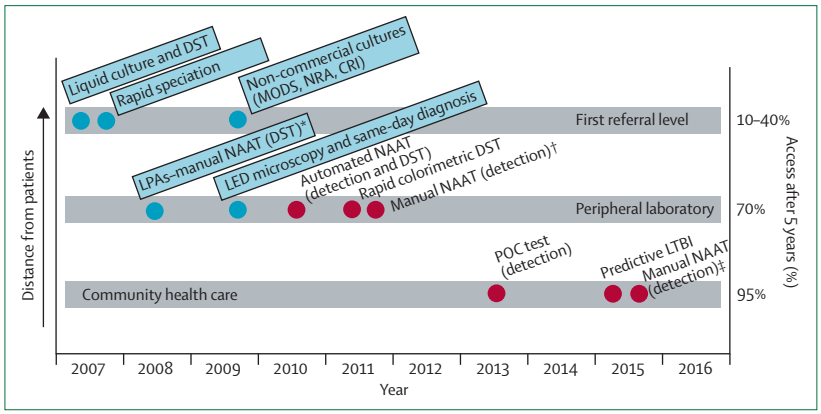


Figure 3: **The tuberculosis diagnostics pipeline**

Technologies in boxes have been endorsed by WHO. DST=drug-susceptibility test. MODS=microscopic observation drug susceptibility. NRA=nitrate reductase assay. CRI=colorimetric redox indicator assay. LPA=line-probe assay. NAAT=nucleic acid amplification test. LED=light-emitting diode. POC=point of care. LTBI=latent tuberculosis infection. *Manual NAAT: technology for *Mycobacterium tuberculosis* drug-susceptibility testing. †Manual NAAT: technology for *M tuberculosis* detection at the peripheral laboratory. ‡Manual NAAT: technology for *M tuberculosis* detection at the community health-care level. Source: adapted from Stop TB Partnership. Global Plan to Stop TB, 2006–2015,[157] and reproduced with permission from author and publisher.

The diagnostics pipeline and new WHO policies

Over the past decade, tangible progress has been made in the development of new tuberculosis diagnostics. The increase in investments has resulted in an expanded pipeline of new diagnostic tests.[162,163] The private sector, led by funding from the Bill & Melinda Gates Foundation, is increasingly engaged in public-private partnerships such as the Foundation for Innovative New Diagnostics (FIND) to develop and deliver a pipeline of tests that are appropriate for disease-endemic countries. Furthermore, under the umbrella of the Global Laboratory Initiative—one of the Working Groups of the Stop TB Partnership—plans are underway for a large scale-up of laboratory services for tuberculosis. For example, UNITAID is providing US$81 million funding for a programme called EXPAND-TB that will supply rapid diagnostics for MDR tuberculosis to 27 high-burden countries.[164] Another example is the allocation of substantial resources to laboratory strengthening by the US President's Emergency Plan for AIDS Relief (PEPFAR).[165]

For the first time in many years, progress is being made in developing a range of diagnostic options for laboratories in disease-endemic countries. The Stop TB Partnership's Retooling Task Force and New Diagnostics Working Group produced a summary on the diagnostics pipeline.[166] Figure 3 shows an updated version of the pipeline,[167] which displays the tests that have been endorsed by WHO between 2007 and 2009. A complete description of existing and novel tuberculosis diagnostics is available elsewhere.[139,168]

Since 2007, several tuberculosis diagnostics have been endorsed by WHO for use in disease-endemic countries (panel 2). In 2007, WHO endorsed the use of liquid culture systems and rapid tests for species confirmation through antigen detection.[169] This WHO policy, along with FIND's negotiations with industry, made implementation of liquid culture systems affordable and feasible for the first time, especially in countries with high HIV prevalence.

Line-probe assays, which are based on reverse hybridisation technology, have consistently shown excellent accuracy for rapid detection of MDR tuberculosis.[175] As a result, in 2008, WHO endorsed the use of these assays for rapid detection of MDR tuberculosis in smear-positive patients.[170] Several non-commercial and less expensive options have been explored for MDR screening of clinical specimens with a variety of culture methods within centralised reference laboratories, including microscopic observation drug susceptibility, thin-layer agar, direct nitrate reductase, and colorimetric redox indicator assays. WHO considered evidence for their accuracy and role, and recommended that selected non-commercial culture and drug-susceptibility testing methods be used as an interim solution in resource-constrained settings, in reference laboratories, or in other laboratories with sufficient culture capacity, while capacity for genotypic or automated liquid

culture and drug-susceptibility testing is being developed.[171] Although non-commercial assays have similar accuracy as do commercial liquid culture systems and cost less, these tests are not standardised and need extensive training, optimisation, and quality assurance before clinical use.

Panel 2: Summary of WHO policies and statements on tuberculosis diagnostics

Liquid media for culture and DST (introduced in 2007)

WHO recommends, as a step-wise approach:

- The use of liquid medium for culture and DST in middle-income and low-income countries.
- Rapid species identification to address the needs for culture and DST, taking into consideration that implementation of liquid systems will be phased, will be integrated into a country-specific comprehensive plan for laboratory-capacity strengthening, and will address several issues including biosafety and training.

Definition of a new sputum-smear-positive tuberculosis case (introduced in 2007)

The revised definition of a new sputum-smear-positive case of pulmonary tuberculosis is based on the presence of at least one acid fast bacilli in at least one sputum sample in countries with a well functioning external quality-assurance system.

Reduction of number of smears for diagnosis of pulmonary tuberculosis (introduced in 2007)

WHO recommends the number of specimens to be examined for screening of tuberculosis cases can be reduced from three to two, in places where a well functioning external quality-assurance system exists, where the workload is very high, and human resources are scarce.

Molecular line-probe assays for rapid screening of patients at risk of MDR tuberculosis (introduced in 2008)

The use of line-probe assays is recommended by WHO, with the following guiding principles:

- Adoption of line-probe assays for rapid detection of MDR tuberculosis should be decided by ministries of health within the context of country plans for appropriate management of patients with MDR tuberculosis, including the development of country-specific screening algorithms and timely access to quality-assured second-line antituberculosis drugs.
- Direct use of line-probe assays on smear-negative clinical specimens is not recommended.
- The use of commercial line-probe assays, rather than in-house assays, is recommended to ensure reliability and reproducibility of results.
- Adoption of line-probe assays does not eliminate the need for conventional culture and DST capability; culture remains necessary for definitive diagnosis of tuberculosis in smear-negative patients, whereas conventional DST is needed to diagnose XDR tuberculosis.

(Continues in next column)

(Continued from previous column)

LED-based microscopy (introduced in 2009–10)

- WHO recommends that conventional fluorescence microscopy be replaced by LED microscopy using auramine staining in all settings where fluorescence microscopy is currently used, and that LED microscopy be phased in as an alternative for conventional Ziehl-Neelsen light microscopy in both high-volume and low-volume laboratories.
- The switch to LED microscopy should be undertaken through a carefully phased implementation plan, with use of LED technologies that meet WHO specifications.

Non-commercial culture DST methods (introduced in 2009–10)

WHO recommends that selected non-commercial culture and DST methods be used as an interim solution in resource-constrained settings, in reference laboratories, or in those with sufficient culture capacity, while capacity for genotypic and/or automated liquid culture and DST are being developed. With due consideration of the above issues, WHO endorses the selective use of one or more of the following non-commercial culture and DST methods:

- Microscopically observed drug susceptibility as direct or indirect tests, for rapid screening of patients suspected of having MDR tuberculosis.
- Nitrate reductase assay, as direct or indirect tests, for screening of patients suspected of having MDR tuberculosis, and acknowledging that time to detection of MDR tuberculosis in indirect application would not be faster than conventional DST methods using solid culture.
- Colorimetric redox indicator methods, as indirect tests on *Mycobacterium tuberculosis* isolates from patients suspected of having MDR tuberculosis, and acknowledging that time to detection of MDR tuberculosis would not be faster (but would be less expensive) than conventional DST methods using commercial liquid culture or molecular line-probe assays.

Same-day diagnosis by microscopy (introduced in 2009–10):

- WHO recommends that countries that have successfully implemented the current WHO policy for a two-specimen case-finding strategy consider a switch to the same-day-diagnosis approach, especially in settings where patients are likely to default from the diagnostic process.
- Countries that are still using the three-specimen case-finding strategy consider a gradual change to the same-day-diagnosis approach, once WHO-recommended external microscopy quality-assurance systems are in place and good quality microscopy results have been documented.
- Changes to a same-day-diagnosis strategy be preceded by a detailed situation assessment of the programmatic, logistical, and operational implications within countries, and supported by a carefully phased implementation plan.

Source: WHO.[169-174] DST=drug-susceptibility testing. MDR=multidrug resistant. XDR=extensively drug resistant. LED=light-emitting diode.

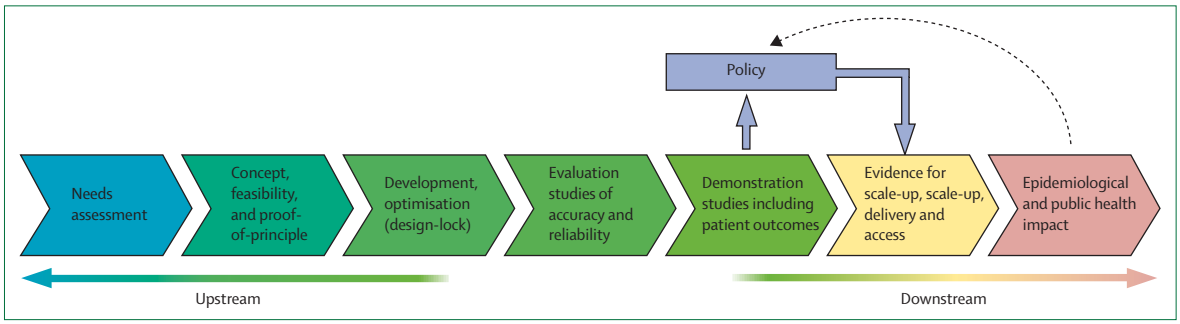


Figure 4: **Schematic showing the pathway to tuberculosis diagnostics, from concept to delivery**
Source: Stop TB Partnership's New Diagnostics Working Group. Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics (2009),[180] and reproduced with permission from author and publisher.

Fluorescence microscopy is widely used in high-income countries since it offers increased sensitivity, and has logistical advantages such as less technician time,[176] but is rarely used in resource-limited countries. Several light-emitting diode (LED) microscopes that can be used in fluorescence microscopy have been developed in the past few years.[177] They are inexpensive, robust, consume little electricity, are highly sensitive, and need less technician time than does Ziehl-Neelsen microscopy. WHO recommended that conventional fluorescence microscopy be replaced by LED microscopy in all settings and that LED microscopy be phased in as an alternative for conventional Ziehl-Neelsen microscopy in both high-volume and low-volume laboratories.[171] Efforts are also underway to minimise diagnostic delays and to improve system efficiency by optimising the number of specimens that are needed and the way in which they are collected (eg, so-called same-day diagnosis, using two sputum smears collected on the same day).[178] In fact, WHO recently endorsed the use of the same-day microscopy approach.[171]

The growing evidence base for tuberculosis diagnostics

Evidence presented to WHO expert committees over the past few years that has informed the endorsement of new technologies (panel 2) included feasibility studies assessing the technology aspect, evaluation studies of the final manufactured product, and large-scale demonstration projects focused on cost, effect, and practicability of use in real-world settings.[179] This extensive and robust platform of evidence is time consuming and expensive to generate, but necessary to lend support to evidence-based policies for tuberculosis diagnosis.[172] An outline for the development of tuberculosis diagnostics, published by the Stop TB Partnership's New Diagnostics Working Group, formalised this evidence platform by describing the development pathway for new tuberculosis diagnostics in detail (figure 4), from initial concept to development, evaluation, delivery, scale-up, and impact assessment.[180]

WHO has assumed a leadership role in ensuring that new tuberculosis diagnostic policies are evidence based,[179] and in line with the grading of recommendations assessment, development and evaluation (GRADE) approach to guideline development.[172] To enable and help with this process, existing systematic reviews on tuberculosis diagnostics, policies, guidelines, and research agendas for diagnosis have been compiled by the Stop TB Partnership's New Diagnostics Working Group.[181] Panel 3 summarises the findings of systematic reviews of various tuberculosis diagnostics.

Optimism for the future

The product pipeline for the future looks promising. In 2009, data were published on the first automated molecular test for tuberculosis, the Xpert MTB/RIF, which was co-developed by the Foundation for Innovative New Diagnostics, Cepheid (Sunnyvale, CA, USA), and the University of Medicine and Dentistry of New Jersey, NJ, USA.[182] This assay, which was CE (Conformité Européenne) marked in 2009, avoids most of the pitfalls of conventional nucleic acid amplification tests (safety, contamination, ease of use, etc), can be done by staff with little training, and can be used for case detection or MDR screening. Data from evaluation trials showed excellent performance in both smear-positive and smear-negative patients, and high accuracy for determination of rifampicin resistance. Thus, this highly sensitive and simple-to-use system can detect *M tuberculosis* directly from sputum in less than 2 h.[182] Data from ongoing demonstration projects are likely to be reviewed by WHO in 2010.

For the diagnosis of latent *M tuberculosis* infection, commercially available IGRAs have emerged as a strong alternative to the tuberculin skin test. These assays have very high specificity and have specific logistical advantages compared with the tuberculin skin test[183] for diagnosis. IGRAs, however, have no role as rule-in tests for active tuberculosis diagnosis for adults in endemic settings.[183] The use of IGRAs is steadily increasing, with several countries with low and intermediate incidence opting to use them, mostly as follow-up tests in people with positive results from tuberculin skin tests, especially in BCG-vaccinated populations.[184] A survey of IGRA guidelines showed much diversity in how various countries recommend and use

IGRAs.[184] The two-step approach (initial tuberculin skin test followed by confirmatory IGRA testing) seems to be the most common strategy, partly because of economic considerations. The optimum strategy for IGRA use is yet to be established.

Although targeted testing and preventive therapy for latent *M tuberculosis* infection is well established in low-incidence countries, the exact role of testing and treatment in disease-endemic countries remains controversial. However, testing for latent *M tuberculosis* infection is receiving increased attention in vulnerable subgroups, such as HIV-infected people and childhood contacts of active tuberculosis cases.[185,186] A WHO policy for IGRAs is under consideration for 2010.

Limitations of the existing diagnostics pipeline

A simple, rapid, inexpensive point-of-care test for active tuberculosis that can perform as well or better than conventional smear microscopy, and can deliver results within minutes without sophisticated equipment or laboratory requirements, is still missing from the development pipeline. Point-of-care diagnostic tests offer important potential advantages for control of diseases such as tuberculosis that need lengthy standardised, decentralised therapy.[142,187,188] Patient, community, and activist groups have urged for increased funding and resources to develop point-of-care tests, and specifications for an ideal point-of-care test have been proposed.[189]

The existing tuberculosis diagnostics pipeline is also restricted with respect to tests that address important diagnostic challenges, especially in HIV-infected people, and children and adults with smear-negative tuberculosis. Unfortunately, most existing tests have shown disappointing performance in smear-negative tuberculosis. Conventional nucleic acid amplification tests have inadequate sensitivity in patients with smear-negative tuberculosis. Improved tests such as Xpert MTB/RIF might have improved sensitivity in these patients, but further validation is needed.[182]

Childhood tuberculosis is a well known diagnostic challenge, and all available tests do poorly in cases of paucibacillary tuberculosis.[190] Furthermore, since young children are unable to produce sputum, alternative specimens such as urine, saliva, or breath condensate would be helpful to use. The absence of a gold standard for childhood tuberculosis and smear-negative tuberculosis is an important impediment to rapid assessment of new diagnostic methods in these high-risk subgroups. One potential solution to the problem of an inadequate gold standard would be to follow up well characterised cohorts of patients after initial testing until tuberculosis is definitely ruled in or out. This type of study could also assess whether use of a new test actually improved patient-important outcomes, rather than examining sensitivity and specificity only.

Although serological antibody tests for tuberculosis have potential as point-of-care tests, their performance

Panel 3: Summary of findings from systematic reviews on tuberculosis diagnostic tests

Diagnosis of active tuberculosis

Sputum-smear microscopy for pulmonary tuberculosis

- FM is on average 10% more sensitive than is conventional microscopy. Specificity of both FM and conventional microscopy is similar. FM is associated with improved time efficiency.
- LED FM performs equivalently to conventional FM, with added benefits of low cost, durability, and ability to use without a darkroom.
- Centrifugation and overnight sedimentation, preceded with any of several chemical methods (including bleach), is slightly more sensitive (6–9%) than is direct microscopy; specificity might be slightly decreased (1–3%) by sputum processing methods.
- When serial sputum specimens are examined, the mean incremental yield and/or increase in sensitivity from examination of third sputum specimen ranges between 2% and 5%.
- A same-day-diagnosis approach (microscopy of two consecutive spot-spot sputum specimens) is equivalent, in terms of diagnostic accuracy, to conventional case-finding strategies by microscopy.

NAATs for pulmonary and extrapulmonary tuberculosis

- NAATs have high specificity and positive predictive value. NAATs, however, have relatively lower (and highly variable) sensitivity and negative predictive value for all forms of tuberculosis, especially in smear-negative and extrapulmonary disease.
- In-house (so-called home brew) NAATs produce highly inconsistent results compared with commercial, standardised NAATs.

Serological antibody detection tests for pulmonary and extrapulmonary tuberculosis

- Commercial serological tests for both pulmonary and extrapulmonary tuberculosis produce highly inconsistent estimates of sensitivity and specificity; none of the assays do well enough to replace microscopy.
- Several potential candidate antigens for inclusion in an antibody detection-based diagnostic test for pulmonary tuberculosis in HIV-infected and uninfected individuals were identified.
- Combinations of select antigens provide higher sensitivities than do single antigens.

ADA for tuberculosis pleuritis, pericarditis, peritonitis

- Measurement of ADA concentrations in pleural, pericardial, and ascitic fluid has high sensitivity and specificity for extrapulmonary tuberculosis.

Interferon γ for tuberculosis pleuritis

- Pleural fluid interferon-γ determination is a sensitive and specific test for the diagnosis of tuberculosis pleuritis.

Phage amplification assays for pulmonary tuberculosis

- Phage-based assays have high specificity but lower and variable sensitivity. Current commercial phage-based assays are limited by high rates of indeterminate results.

Automated liquid cultures for pulmonary tuberculosis

- Automated liquid cultures are more sensitive than are solid cultures; time to detection is more rapid than for solid cultures.

Diagnosis of latent tuberculosis

TST for latent tuberculosis infection

- Individuals who had received BCG vaccination are more likely to have a positive TST; the effect of BCG on TST results is less after 15 years; positive TST with indurations of >15 mm are more likely to be the result of tuberculosis infection than of BCG vaccination.

(Continues on next page)

(Continued from previous page)

- The effect on TST of BCG received in infancy is small, especially 10 years after vaccination. BCG received after infancy produces more frequent, more persistent, and larger TST reactions. NTM infection is not a clinically important cause of false-positive TST, apart from in populations with a high prevalence of NTM sensitisation and a very low prevalence of tuberculosis infection.

T-cell-based IGRAs for latent tuberculosis infection

- IGRAs have excellent specificity (higher than the TST), and are unaffected by previous BCG vaccination.
- IGRAs cannot distinguish between latent tuberculosis infection and active tuberculosis, and have no role for active tuberculosis diagnosis in adults.
- IGRAs correlate well with markers of tuberculosis exposure in low-incidence countries.
- IGRA sensitivity varies across populations and tends to be lower in high-endemic countries and in HIV-infected individuals.

Diagnosis of drug-resistant tuberculosis

Phage amplification assays for rapid detection of rifampicin resistance

- Commercial phage amplification assays produce variable results when used directly on sputum specimens.
- Studies have raised concerns about contamination, false positive results, and technical assay failures.

Line-probe assays: INNO-LiPA Rif and GenoType MTBDR assays for rapid detection of rifampicin resistance

- The INNO-LiPA Rif assay is a highly sensitive and specific test for the detection of rifampicin resistance in culture isolates. The test has lower sensitivity when used directly on clinical specimens.
- GenoType MTBDR assays have excellent sensitivity and specificity for rifampicin resistance, even when directly used on clinical specimens.

CRI methods and NRA for rapid detection of rifampicin and isoniazid resistance

- Colorimetric methods are sensitive and specific for the detection of rifampicin and isoniazid resistance in culture isolates. CRIs use inexpensive non-commercial supplies and equipment and have a rapid turnaround time (7 days).
- NRA has high accuracy when used to detect rifampicin and isoniazid resistance in culture isolates. Data for its use when directly applied to clinical specimens are scarce, but results are promising. NRA is simple, uses inexpensive non-commercial supplies and equipment, and has a rapid turnaround time (7–14 days) compared with conventional methods.

MODS for rapid detection of rifampicin and isoniazid resistance

- MODS has high accuracy when testing for rifampicin resistance, but shows slightly lower sensitivity when detecting isoniazid resistance.
- MODS seems to do equally well with use of direct patient specimens and culture isolates.
- MODS uses non-commercial supplies and equipment, and has a rapid turnaround time (10 days) compared with conventional methods.

TLA for rapid detection of rifampicin and isoniazid resistance

- Data assessing TLA for the detection of drug susceptibility are scarce; however, all studies so far have reported 100% concordance with their reference standards.
- TLA uses inexpensive non-commercial supplies and equipment, and has a rapid turnaround time (11 days) compared with conventional methods.

FM=fluorescence microscopy. LED=light-emitting diode. NAATs=nucleic acid amplification tests. ADA=adenosine deaminase. TST=tuberculin skin test. NTM=non-tuberculous mycobacterial. IGRAs=interferon-γ release assays. CRI=colorimetric redox indicator. NRA=nitrate reductase assays. MODS=microscopically observed drug susceptibility. TLA=thin-layer agar. Adapted from reference 172 and reproduced with permission from the author, under creative commons open access license.

thus far has been disappointing.[191,192] Urine mycobacterial antigen (eg, lipoarabinomannan) detection tests are attractive options for point-of-care testing. Results of studies of detection of urinary lipoarabinomannan have been variable but generally suboptimum, although somewhat better in patients with advanced HIV infection.[193–197] Alternative detection targets are being sought, such as *M tuberculosis* tr-DNA.[198]

Several options are being explored for simpler, less expensive point-of-care and multiplexed assay formats in the future, including manual molecular testing that can be done in peripheral settings, lab-on-chip approaches that can be used to detect several infections simultaneously, antigen detection on highly sensitive platforms, and antibody detection with panels of recently identified antigens of diagnostic value. Owing to growing interest and funding for new methods and biomarkers, several agencies, industries, and groups are working on developing point-of-care platforms for tuberculosis, including novel serological assays, detection of volatile organic compounds in breath, handheld molecular devices, microchip technologies, and tests that exploit approaches such as microfluidics, nanotechnology, proteomics, and metabolomics.[139,162,168,187]

Overcoming barriers for implementation in tuberculosis control programmes

What will be the outcome of all this technology development for tuberculosis diagnostics, and how can this progress be translated into concrete gains in control of tuberculosis? The effect of new tests will depend largely on the extent of their introduction into the global public sector, which will itself depend partly on policy decisions made by international technical agencies such as WHO, and by donors, and ultimately by national tuberculosis programmes in countries of low and middle income. So far, most evaluations of diagnostic methods have reported only sensitivity and specificity; many of these studies were poorly designed and incompletely reported.[172,199] Some have assessed time-to-test result, and a few have reported unit costs. However, to lend support to the introduction of new diagnostic technologies, broader evidence is needed, including implementation issues.[180] For example, the performance of new tests in programmatic conditions should be studied; tests done by experts in carefully controlled research settings are not likely to be indicative of future field performance. In addition to unit costs, costing studies should include costs for labour, equipment depreciation, initial and ongoing training, supervision, and quality control.[200]

Future studies of completed diagnostic products have to go beyond test accuracy and aim to generate evidence for the incremental value of new tests, their effect on patient outcomes, and their use for diagnostic decision making and cost-effectiveness.[172,199] Operational research is also essential to improve service delivery and to understand why diagnosis is delayed or missed, and to guide optimum implementation of new methods. To help with these types

of research, the TB Research Movement—recently initiated by the Stop TB Partnership and WHO—is engaging tuberculosis researchers, tuberculosis programme managers, and affected communities in a collaborative and concerted strategic effort to increase the scope, scale, and speed of tuberculosis research across the continuum, linking together basic research, development of new methods, and operational research.[201]

The GRADE approach,[202] now being used by WHO, was originally designed for interventions such as drugs and vaccines, for which the product is the health intervention itself and the use of the product can be largely judged by its safety and effectiveness alone. Diagnostics, however, are only the start of a health intervention, and their effect will depend on where and how they are used, and what clinical decisions they can lend support to. The GRADE approach has been adapted and applied to diagnostic tests,[203,204] but will need to be further adapted or supplemented by careful considerations of the diversity and challenges of health systems when examining diagnostics aimed at public-sector populations in developing countries. Although GRADE has its limitations and can be improved and adapted for tuberculosis diagnostics, it is a major advance compared with the conventional policy-making process.[205]

Inadequate funding is another major barrier that needs to be overcome. New tuberculosis diagnostics will be of no practical value if they are not readily available at points of care in endemic areas, and if they are not taken seriously by governments of developing countries. Insufficient commitment to tuberculosis control by many developing country governments is largely responsible for poor programme performance. The Global Plan to Stop TB, 2006–2015, estimated that at least US$9 billion ($900 million per year) should be spent on tuberculosis research and development between 2006 and 2015 to develop new drugs, diagnostics, and vaccines.[167] The budget needed for tuberculosis diagnostics was $516 million; yet according to the 2009 Treatment Action Group and Stop TB Partnership reports, development for tuberculosis diagnostics received only $50 million in 2008.[206,207] This amount represented only 10% of the total funding for tuberculosis research and development.[207] Furthermore, philanthropic grants are outstripping government funding for tuberculosis research.[207]

To overcome this worrisome trend in declining public-sector investment, governments in all countries, especially industrialised countries, need to increase their funding for tuberculosis research and development.[167,206,207] Emerging and rapidly growing economies such as China, India, Brazil, and South Africa can and should increase their investments in tuberculosis, especially since these countries account for a large proportion of the global tuberculosis burden. Countries such as China and India can also make a big contribution by producing locally manufactured, low-cost generic tuberculosis drugs, diagnostics, and vaccines. In the long term, these countries have the potential to spearhead the next wave of innovation in tuberculosis research and development.

Conclusions

The need for a more accurate, inexpensive point-of-care tuberculosis diagnostic test that is applicable in tuberculosis and HIV endemic areas is greater nowadays than ever before, and will be crucial for achieving global tuberculosis control. Several modelling studies[208–215] suggest that new diagnostics for tuberculosis disease and MDR tuberculosis could have an important effect within populations, especially in disease-endemic countries, although improving population health and health services, and economic growth, might be as important.[216,217] Clinical and field studies are needed to assess whether programmatic introduction of new diagnostics contributes to improved individual patient outcomes and a measurable beneficial public health effect. After nearly a century of neglect and underinvestment, the tuberculosis diagnostics pipeline has rapidly grown, with several technologies showing great promise. Indeed, several have already been endorsed by WHO and are being introduced into clinical use. This progress needs to be translated into improving the lives of patients with tuberculosis, and reducing the future incidence of tuberculosis. This aim can and must be achieved, but will need strong political commitment, sustained funding, and engagement of public and private stakeholders and civil society. Donors and governments have to synergise their activities to ensure maximum programme performance for optimum care for patients with tuberculosis and with both tuberculosis and HIV infection.

To advance the area of tuberculosis biomarkers to that needed for registration, substantial investments will be required to undertake the necessary studies. Studies of MDR tuberculosis have been advocated by some as a rich source of poor outcomes for biomarkers research. However, whether markers predicting failure due to resistance will necessarily also predict relapse (which seems somewhat paradoxically to occur infrequently in MDR tuberculosis) is uncertain.[218] As studies are undertaken to shorten MDR treatment, we might need to rely on biomarkers for relapse developed in drug-sensitive disease to guide them. Tuberculosis incidence rarely approaches 1% in the general population even in high-prevalence countries, hampering prospective studies. Ethical concerns preclude natural history studies in high-risk patients, such as children or people with HIV infection, meaning that isoniazid preventive therapy should be offered. As a result, studies to validate markers that predict the transition from health to illness (or vice versa) in these populations will necessarily be large and protracted. If the plethora of potential biomarkers described here is to be converted into clinically useful tests, we not only need continuing research, but also

improved funding to synergise and improve multidisciplinary cross-cutting collaborations between scientists working with cohorts of patients and contacts participating in clinical trials of new drug regimens, diagnostics, and vaccines.

Contributors

AZ conceived the article outline and selected and assigned authors' roles. The literature search and the first and subsequent drafts of the report were developed by RSW (biomarkers) and MP (diagnostics). TMD, GW, and AZ contributed to writing of the biomarkers section, and MDP, DM, and AZ contributed to writing of the diagnostics section. AZ merged the two sections as the final article with contributions from all authors. All authors read and approved the final versions of the two sections before submission.

Steering committee

This article is part of *The Lancet* Series on tuberculosis, which was developed and coordinated by Alimuddin Zumla (University College London Medical School, London, UK); Mario C Raviglione (Stop TB Department, WHO, Geneva, Switzerland); and Ben Marais (University of Stellenbosch, Stellenbosch, South Africa).

Conflicts of interests

RSW is employed by Pfizer, USA. AZ is principal investigator of the EuropeAID Active Detection of Active Tuberculosis (ADAT) and European Union Framework 7 Trans-renal DNA (EU-FW7-TrDNA) projects, which are assessing new tuberculosis diagnostics, and serves on the Stop TB Research Movement Task Force. MDP is the Chief Scientific Officer of Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland, a non-profit agency that works with several industry partners in developing and evaluating new diagnostics for neglected infectious diseases. MP serves as an external consultant for FIND; serves as a co-chair of the Stop TB Partnership's New Diagnostics Working Group; and serves as chair of the Task Force of the Stop TB Research Movement. GW receives support from GlaxoSmithKline, Bill & Melinda Gates Foundation, Aeras Foundation, TB Alliance, and EDCTP. TMD receives support from EDCTP and EU-FP7. None of the funding agencies had any role in the development or submission of this report. DM declares that he has no conflicts of interest.

Acknowledgments

AZ receives support from the EU-FP7, EDCTP, Global Alliance for TB Drug development, EuropeAID-ADAT, UK Medical Research Council, and the UK NIHR CBRC. TMD receives support from the EU-FP7; EDCTP; Research Council of Norway; Netherlands-African partnership for capacity development and clinical interventions against poverty-related diseases; and the Danish Agency for Science, Technology and Innovation. MP is a recipient of a New Investigator Award from the Canadian Institutes of Health Research (CIHR), and receives support from EDCTP and European Commission (TBSusgent, EU-FP7). DM is recipient of a career award from the Fonds de la Recherche en Santé du Québec (FRSQ). None of the funding agencies had any role in the development or submission of this report.

References

1 Biomarkers working group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001; **69:** 89–95.

2 Mitchison DA. Assessment of new sterilizing drugs for treating pulmonary tuberculosis by culture at 2 months. *Am Rev Respir Dis* 1993; **147:** 1062–63.

3 The Tuberculosis Trials Consortium. Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial. *Lancet* 2002; **360:** 528–34.

4 Wallis RS, Wang C, Doherty TM, et al. Biomarkers for tuberculosis disease activity, cure, and relapse. *Lancet Infect Dis* 2010; **10:** 68–69.

5 Rustomjee R, Diacon AH, Allen J, et al. Early bactericidal activity and pharmacokinetics of the Diarylquinoline TMC 207 in pulmonary tuberculosis. *Antimicrob Agents Chemother* 2008; **52:** 2831–35.

6 Davies GR, Brindle R, Khoo SH, Aarons LJ. Use of nonlinear mixed-effects analysis for improved precision of early pharmacodynamic measures in tuberculosis treatment. *Antimicrob Agents Chemother* 2006; **50:** 3154–56.

7 Epstein MD, Schluger NW, Davidow AL, Bonk S, Rom WN, Hanna B. Time to detection of Mtb tuberculosis in sputum culture correlates with outcome in patients receiving treatment for pulmonary tuberculosis. *Chest* 1998; **113:** 379–86.

8 Wallis RS, Perkins M, Phillips M, et al. Induction of the antigen 85 complex of *M. tuberculosis* in sputum: a determinant of outcome in pulmonary tuberculosis. *J Infect Dis* 1998; **178:** 1115–21.

9 Rustomjee R, Lienhardt C, Kanyok T, et al. A Phase II study of the sterilising activities of ofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 2008; **12:** 128–38.

10 Wallis RS, Perkins M, Phillips M, et al. Predicting the outcome of therapy for pulmonary tuberculosis. *Am J Respir Crit Care Med* 2000; **161:** 1076–80.

11 Wallis RS, Phillips M, Johnson JL, et al. Inhibition of INH-induced expression of *M. tuberculosis* antigen 85 in sputum: a potential surrogate marker in TB chemotherapy trials. *Antimicrob Agents Chemother* 2001; **45:** 1302–04.

12 Desjardin LE, Perkins MD, Wolski K, et al. Measurement of sputum Mtb tuberculosis messenger RNA as a surrogate for response to chemotherapy. *Am J Respir Crit Care Med* 1999; **160:** 203–10.

13 Li L, Mahan CS, Palaci M, et al. Sputum Mycobacterium tuberculosis mRNA as a marker of bacteriologic clearance in response to anti-tuberculosis therapy. *J Clin Microbiol* 2010; **48:** 46–51.

14 Garton NJ, Waddell SJ, Sherratt AL, et al. Cytological and transcript analyses reveal fat and lazy persister-like bacilli in tuberculous sputum. *PLoS Med* 2008; **5:** e75.

15 Cannas A, Goletti D, Girardi E, et al. Mycobacterium tuberculosis DNA detection in soluble fraction of urine from pulmonary tuberculosis patients. *Int J Tuberc Lung Dis* 2008; **12:** 146–51.

16 Cannas A, Calvo L, Chiacchio T, et al. *Mycobacterium tuberculosis* DNA detection in the urine from pulmonary tuberculosis patients (abstract). Seventh International Conference on the Pathogenesis of Mycobacterial Infections; June, 2008; Saltsjobaden, Sweden (abstr P17: 39).

17 Shah M, Variava E, Holmes CB, et al. Diagnostic accuracy of a urine lipoarabinomannan test for tuberculosis in hospitalized patients in a high HIV prevalence setting. *J Acquir Immune Defic Syndr* 2009; **52:** 145–51.

18 Mutetwa R, Boehme C, Dimairo M, et al. Diagnostic accuracy of commercial urinary lipoarabinomannan detection in African tuberculosis suspects and patients. *Int J Tuberc Lung Dis* 2009; **13:** 1253–59.

19 Reither K, Saathoff E, Jung J, et al. Low sensitivity of a urine LAM-ELISA in the diagnosis of pulmonary tuberculosis. *BMC Infect Dis* 2009; **9:** 141.

20 Daley P, Michael JS, Hmar P, et al. Blinded evaluation of commercial urinary lipoarabinomannan for active tuberculosis: a pilot study. *Int J Tuberc Lung Dis* 2009; **13:** 989–95.

21 Boehme C, Molokova E, Minja F, et al. Detection of mycobacterial lipoarabinomannan with an antigen-capture ELISA in unprocessed urine of Tanzanian patients with suspected tuberculosis. *Trans R Soc Trop Med Hyg* 2005; **99:** 893–900.

22 Tessema TA, Bjune G, Assefa G, Svenson S, Hamasur B, Bjorvatn B. Clinical and radiological features in relation to urinary excretion of lipoarabinomannan in Ethiopian tuberculosis patients. *Scand J Infect Dis* 2002; **34:** 167–71.

23 Choudhry V, Saxena RK. Detection of Mycobacterium tuberculosis antigens in urinary proteins of tuberculosis patients. *Eur J Clin Microbiol Infect Dis* 2002; **21:** 1–5.

24 Singh KK, Dong Y, Hinds L, et al. Combined use of serum and urinary antibody for diagnosis of tuberculosis. *J Infect Dis* 2003; **188:** 371–77.

25 Kashino SS, Pollock N, Napolitano DR, Rodrigues V Jr, Campos-Neto A. Identification and characterization of Mycobacterium tuberculosis antigens in urine of patients with active pulmonary tuberculosis: an innovative and alternative approach of antigen discovery of useful microbial molecules. *Clin Exp Immunol* 2008; **153:** 56–62.

26 Napolitano DR, Pollock N, Kashino SS, Rodrigues V Jr, Campos-Neto A. Identification of Mycobacterium tuberculosis ornithine carboamyltransferase in urine as a possible molecular marker of active pulmonary tuberculosis. *Clin Vaccine Immunol* 2008; **15:** 638–43.

27 Syhre M, Chambers ST. The scent of Mycobacterium tuberculosis. *Tuberculosis (Edinb)* 2008; **88:** 317–23.

28 Phillips M, Cataneo RN, Condos R, et al. Volatile biomarkers of pulmonary tuberculosis in the breath. *Tuberculosis (Edinb)* 2007; **87:** 44–52.

29 Wallis RS, Palaci M, Vinhas S, et al. A whole blood bactericidal assay for tuberculosis. *J Infect Dis* 2001; **183:** 1300–03.

30 Janulionis E, Sofer C, Song HY, Wallis RS. Lack of activity of oral clofazimine against intracellular M. tuberculosis in whole blood culture. *Antimicrob Agents Chemother* 2004; **48:** 3133–35.

31 Wallis RS, Vinhas SA, Johnson JL, et al. Whole blood bactericidal activity during treatment of pulmonary tuberculosis. *J Infect Dis* 2003; **187:** 270–78.

32 Veenstra H, Crous I, Brahmbhatt S, et al. Changes in the kinetics of intracellular IFN-gamma production in TB patients during treatment. *Clin Immunol* 2007; **124:** 336–44.

33 Carrara S, Vincenti D, Petrosillo N, Amicosante M, Girardi E, Goletti D. Use of a T cell-based assay for monitoring efficacy of antituberculosis therapy. *Clin Infect Dis* 2004; **38:** 754–56.

34 Sauzullo I, Mengoni F, Lichtner M, et al. In vivo and in vitro effects of antituberculosis treatment on mycobacterial interferon-gamma T cell response. *PLoS One* 2009; **4:** e5187.

35 Aiken AM, Hill PC, Fox A, et al. Reversion of the ELISPOT test after treatment in Gambian tuberculosis cases. *BMC Infect Dis* 2006; **6:** 66.

36 Kobashi Y, Mouri K, Yagi S, Obase Y, Miyashita N, Oka M. Transitional changes in T-cell responses to Mycobacterium tuberculosis-specific antigens during treatment. *J Infect* 2009; **58:** 197–204.

37 Ribeiro S, Dooley K, Hackman J, et al. T-SPOT.TB responses during treatment of pulmonary tuberculosis. *BMC Infect Dis* 2009; **9:** 23.

38 Wassie L, Demissie A, Aseffa A, et al. Ex vivo cytokine mRNA levels correlate with changing clinical status of ethiopian TB patients and their contacts over time. *PLoS One* 2008; **3:** e1522.

39 Dheda K, Chang JS, Breen RA, et al. In vivo and in vitro studies of a novel cytokine, interleukin 4delta2, in pulmonary tuberculosis. *Am J Respir Crit Care Med* 2005; **172:** 501–08.

40 Siawaya JF, Bapela NB, Ronacher K, Beyers N, van Helden P, Walzl G. Differential expression of interleukin-4 (IL-4) and IL-4 delta 2 mRNA, but not transforming growth factor beta (TGF-beta), TGF-beta RII, Foxp3, gamma interferon, T-bet, or GATA-3 mRNA, in patients with fast and slow responses to antituberculosis treatment. *Clin Vaccine Immunol* 2008; **15:** 1165–70.

41 Wallis RS, Helfand MS, Whalen C, et al. Immune activation, allergic drug toxicity, and mortality in HIV-positive tuberculosis. *Tuber Lung Dis* 1996; **77:** 516–23.

42 Immanuel C, Rajeswari R, Rahman F, Kumaran PP, Chandrasekaran V, Swamy R. Serial evaluation of serum neopterin in HIV seronegative patients treated for tuberculosis. *Int J Tuberc Lung Dis* 2001; **5:** 185–90.

43 Turgut T, Akbulut H, Deveci F, Kacar C, Muz MH. Serum interleukin-2 and neopterin levels as useful markers for treatment of active pulmonary tuberculosis. *Tohoku J Exp Med* 2006; **209:** 321–28.

44 Hosp M, Elliott AM, Raynes JG, et al. Neopterin, beta 2-microglobulin, and acute phase proteins in HIV-1-seropositive and -seronegative Zambian patients with tuberculosis. *Lung* 1997; **175:** 265–75.

45 Fuchs D, Hausen A, Kofler M, Kosanowski H, Reibnegger G, Wachter H. Neopterin as an index of immune response in patients with tuberculosis. *Lung* 1984; **162:** 337–46.

46 Baylan O, Balkan A, Inal A, et al. The predictive value of serum procalcitonin levels in adult patients with active pulmonary tuberculosis. *Jpn J Infect Dis* 2006; **59:** 164–67.

47 Nyamande K, Lalloo UG. Serum procalcitonin distinguishes CAP due to bacteria, Mycobacterium tuberculosis and PJP. *Int J Tuberc Lung Dis* 2006; **10:** 510–15.

48 Polzin A, Pletz M, Erbes R, et al. Procalcitonin as a diagnostic tool in lower respiratory tract infections and tuberculosis. *Eur Respir J* 2003; **21:** 939–43.

49 Prat C, Dominguez J, Andreo F, et al. Procalcitonin and neopterin correlation with aetiology and severity of pneumonia. *J Infect* 2006; **52:** 169–77.

50 Schleicher GK, Herbert V, Brink A, et al. Procalcitonin and C-reactive protein levels in HIV-positive subjects with tuberculosis and pneumonia. *Eur Respir J* 2005; **25:** 688–92.

51 Kandemir O, Uluba B, Polat G, Sezer C, Camdeviren H, Kaya A. Elevation of procalcitonin level in patients with pulmonary tuberculosis and in medical staff with close patient contact. *Arch Med Res* 2003; **34:** 311–14.

52 Wallis RS, van Vuuren C, Potgieter S. Adalimumab treatment of life-threatening tuberculosis. *Clin Infect Dis* 2009; **48:** 1429–32.

53 Tang BM, Eslick GD, Craig JC, McLean AS. Accuracy of procalcitonin for sepsis diagnosis in critically ill patients: systematic review and meta-analysis. *Lancet Infect Dis* 2007; **7:** 210–17.

54 Lawn SD, Wiktor S, Coulibaly D, Ackah AN, Lal RB. Serum C-reactive protein and detection of tuberculosis in persons co-infected with the human immunodeficiency virus. *Trans R Soc Trop Med Hyg* 2001; **95:** 41–42.

55 Bajaj G, Rattan A, Ahmad P. Prognostic value of 'C' reactive protein in tuberculosis. *Indian Pediatr* 1989; **26:** 1010–13.

56 Scott GM, Murphy PG, Gemidjioglu ME. Predicting deterioration of treated tuberculosis by corticosteroid reserve and C-reactive protein. *J Infect* 1990; **21:** 61–69.

57 Plit ML, Theron AJ, Fickl H, Van Rensburg CE, Pendel S, Anderson R. Influence of antimicrobial chemotherapy and smoking status on the plasma concentrations of vitamin C, vitamin E, beta-carotene, acute phase reactants, iron and lipid peroxides in patients with pulmonary tuberculosis. *Int J Tuberc Lung Dis* 1998; **2:** 590–96.

58 Lee JH, Chang JH. Changes of plasma interleukin-1 receptor antagonist, interleukin-8 and other serologic markers during chemotherapy in patients with active pulmonary tuberculosis. *Korean J Intern Med* 2003; **18:** 138–45.

59 Baynes R, Bezwoda W, Bothwell T, Khan Q, Mansoor N. The non-immune inflammatory response: serial changes in plasma iron, iron-binding capacity, lactoferrin, ferritin and C-reactive protein. *Scand J Clin Lab Invest* 1986; **46:** 695–704.

60 Walzl G, Ronacher K, Djoba Siawaya JF, Dockrell HM. Biomarkers for TB treatment response: Challenges and future strategies. *J Infect* 2008; **57:** 103–09.

61 Lai CK, Wong KC, Chan CH, et al. Circulating adhesion molecules in tuberculosis. *Clin Exp Immunol* 1993; **94:** 522–26.

62 Demir T, Yalcinoz C, Keskinel I, Demiroz F, Yildirim N. sICAM-1 as a serum marker in the diagnosis and follow-up of treatment of pulmonary tuberculosis. *Int J Tuberc Lung Dis* 2002; **6:** 155–59.

63 Mukae H, Ashitani J, Tokojima M, Ihi T, Kohno S, Matsukura S. Elevated levels of circulating adhesion molecules in patients with active pulmonary tuberculosis. *Respirology* 2003; **8:** 326–31.

64 Djoba Siawaya JF, Bapela NB, Ronacher K, et al. Immune parameters as markers of tuberculosis extent of disease and early prediction of anti-tuberculosis chemotherapy response. *J Infect* 2008; **56:** 340–47.

65 Eugen-Olsen J, Gustafson P, Sidenius N, et al. The serum level of soluble urokinase receptor is elevated in tuberculosis patients and predicts mortality during treatment: a community study from Guinea-Bissau. *Int J Tuberc Lung Dis* 2002; **6:** 686–92.

66 Rosas-Taraco AG, Salinas-Carmona MC, Revol A, Rendon A, Caballero-Olin G, Arce-Mendoza AY. Expression of CDllc in blood monocytes as biomarker for favorable response to antituberculosis treatment. *Arch Med Res* 2009; **40:** 128–31.

67 Agranoff D, Fernandez-Reyes D, Papadopoulos MC, et al. Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. *Lancet* 2006; **368:** 1012–21.

68 Brahmbhatt S, Black GF, Carroll NM, et al. Immune markers measured before treatment predict outcome of intensive phase tuberculosis therapy. *Clin Exp Immunol* 2006; **146:** 243–52.

69 Mistry R, Cliff JM, Clayton C, et al. Gene expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. *J Infect Dis* 2007; **195:** 357–65.

70 Jacobsen M, Repsilber D, Gutschmidt A, et al. Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. *J Mol Med* 2007; **85:** 613–21.

71 Lalvani A, Pathan AA, Durkan H, et al. Enhanced contact tracing and spatial tracking of Mycobacterium tuberculosis infection by enumeration of antigen-specific T cells. *Lancet* 2001; **357:** 2017–21.

72 Jackson-Sillah D, Hill PC, Fox A, et al. Screening for tuberculosis among 2381 household contacts of sputum-smear-positive cases in The Gambia. *Trans R Soc Trop Med Hyg* 2007; **101:** 594–601.

73 Bakir M, Millington KA, Soysal A, et al. Prognostic value of a T-cell-based, interferon-gamma biomarker in children with tuberculosis contact. *Ann Intern Med* 2008; **149:** 777–87.

74 Aichelburg MC, Rieger A, Breitenecker F, et al. Detection and prediction of active tuberculosis disease by a whole-blood interferon-gamma release assay in HIV-1-infected individuals. *Clin Infect Dis* 2009; **48:** 954–62.

75 Doherty TM, Demissie A, Olobo J, et al. Immune responses to the Mycobacterium tuberculosis-specific antigen ESAT-6 signal subclinical infection among contacts of tuberculosis patients. *J Clin Microbiol* 2002; **40:** 704–06.

76 Diel R, Loddenkemper R, Meywald-Walter K, Niemann S, Nienhaus A. Predictive value of a whole blood IFN-gamma assay for the development of active tuberculosis disease after recent infection with Mycobacterium tuberculosis. *Am J Respir Crit Care Med* 2008; **177:** 1164–70.

77 Higuchi K, Harada N, Fukazawa K, Mori T. Relationship between whole-blood interferon-gamma responses and the risk of active tuberculosis. *Tuberculosis (Edinb)* 2008; **88:** 244–48.

78 Petrucci R, Abu AN, Gurgel RQ, et al. Interferon gamma, interferon-gamma-induced-protein 10, and tuberculin responses of children at high risk of tuberculosis infection. *Pediatr Infect Dis J* 2008; **27:** 1073–77.

79 Whittaker E, Gordon A, Kampmann B. Is IP-10 a better biomarker for active and latent tuberculosis in children than IFNgamma? *PLoS ONE* 2008; **3:** e3901.

80 Ruhwald M, Bjerregaard-Andersen M, Rabna P, Eugen-Olsen J, Ravn P. IP-10, MCP-1, MCP-2, MCP-3, and IL-1RA hold promise as biomarkers for infection with M. tuberculosis in a whole blood based T-cell assay. *BMC Res Notes* 2009; **2:** 19.

81 Demissie A, Wassie L, Abebe M, et al. The 6-kilodalton early secreted antigenic target-responsive, asymptomatic contacts of tuberculosis patients express elevated levels of interleukin-4 and reduced levels of gamma interferon. *Infect Immun* 2006; **74:** 2817–22.

82 Demissie A, Abebe M, Aseffa A, et al. Healthy individuals that control a latent infection with Mycobacterium tuberculosis express high levels of Th1 cytokines and the IL-4 antagonist IL-4delta2. *J Immunol* 2004; **172:** 6938–43.

83 Weldingh K, Andersen P. ESAT-6/CFP10 skin test predicts disease in M. tuberculosis-infected guinea pigs. *PLoS One* 2008; **3:** e1978.

84 Edwards LB, Acquaviva FA, Livesay VT. Identification of tuberculous infected: dual tests and density of reaction. *Am Rev Respir Dis* 1973; **108:** 1334–39.

85 Ozdemir D, Cesur S, Annakkaya AN, et al. Serum neopterin concentrations in healthy healthcare workers compared with healthy controls and patients with pulmonary tuberculosis. *Med Sci Monit* 2006; **12:** CR521–24.

86 Hoft DF, Blazevic A, Abate G, et al. A new recombinant bacille Calmette-Guerin vaccine safely induces significantly enhanced tuberculosis-specific immunity in human volunteers. *J Infect Dis* 2008; **198:** 1491–501.

87 Soares AP, Scriba TJ, Joseph S, et al. Bacillus Calmette-Guerin vaccination of human newborns induces T cells with complex cytokine and phenotypic profiles. *J Immunol* 2008; **180:** 3569–77.

88 Hawkridge T, Scriba TJ, Gelderbloem S, et al. Safety and Immunogenicity of a New Tuberculosis Vaccine, MVA85A, in Healthy Adults in South Africa. *J Infect Dis* 2008; **198:** 544–52.

89 Davids V, Hanekom WA, Mansoor N, et al. The effect of bacille calmette-guerin vaccine strain and route of administration on induced immune responses in vaccinated infants. *J Infect Dis* 2006; **193:** 531–36.

90 Hanekom WA, Hughes J, Mavinkurve M, et al. Novel application of a whole blood intracellular cytokine detection assay to quantitate specific T-cell frequency in field studies. *J Immunol Methods* 2004; **291:** 185–95.

91 Kampmann B, Gaora PO, Snewin VA, Gares MP, Young DB, Levin M. Evaluation of human antimycobacterial immunity using recombinant reporter mycobacteria. *J Infect Dis* 2000; **182:** 895–901.

92 Cheon SH, Kampmann B, Hise AG, et al. Bactericidal activity in whole blood as a potential surrogate marker of immunity after vaccination against tuberculosis. *Clin Diagn Lab Immunol* 2002; **9:** 901–07.

93 Kampmann B, Tena GN, Mazazi S, Young D, Eley B, Levin M. A novel human in vitro system to evaluate antimycobacterial vaccines. *Infect Immun* 2004; **72:** 6401–07.

94 Tena GN, Young DB, Eley B, et al. Failure to control growth of mycobacteria in blood from children infected with human immunodeficiency virus, and its relationship to T cell function. *J Infect Dis* 2003; **187:** 1544–51.

95 Kampmann B, Tena-Coki GN, Nicol M, Levin M, Eley B. Reconstitution of antimycobacterial immune responses in HIV-infected children receiving HAART. *AIDS* 2006; **20:** 1011–18.

96 Martineau AR, Wilkinson RJ, Wilkinson KA, et al. A single dose of vitamin D enhances immunity to mycobacteria. *Am J Respir Crit Care Med* 2007; **176:** 208–13.

97 Saliu O, Sofer C, Stein DS, Schwander SK, Wallis RS. Tumor necrosis factor blockers: differential effects on mycobacterial immunity. *J Infect Dis* 2006; **194:** 486–92.

98 Hoft DF, Worku S, Kampmann B, et al. Investigation of the relationships between immune-mediated inhibition of mycobacterial growth and other potential surrogate markers of protective mycobacterium tuberculosis immunity. *J Infect Dis* 2002; **186:** 1448–57.

99 Wallis RS, Vinhas S, Janulionis E. Strain specificity of antimycobacterial immunity in whole blood culture after cure of tuberculosis. *Tuberculosis (Edinb)* 2009; **89:** 221–24.

100 Canaday DH, Wilkinson RJ, Li Q, Harding CV, Silver RF, Boom WH. CD4(+) and CD8(+) T cells kill intracellular Mycobacterium tuberculosis by a perforin and Fas/Fas ligand-independent mechanism. *J Immunol* 2001; **167:** 2734–42.

101 Cheng SH, Walker L, Poole J, et al. Demonstration of increased anti-mycobacterial activity in peripheral blood monocytes after BCG vaccination in British school children. *Clin Exp Immunol* 1988; **74:** 20–25.

102 Conde MB, Efron A, Loredo C, et al. Moxifloxacin versus ethambutol in the initial treatment of tuberculosis: a double-blind, randomised, controlled phase II trial. *Lancet* 2009; **373:** 1183–89.

103 Dorman SE, Johnson JL, Goldberg S, et al. Substitution of moxifloxacin for isoniazid during intensive phase treatment of pulmonary tuberculosis. *Am J Respir Crit Care Med* 2009; **180:** 273–80.

104 Burman WJ, Goldberg S, Johnson JL, et al. Moxifloxacin versus ethambutol in the first 2 months of treatment for pulmonary tuberculosis. *Am J Respir Crit Care Med* 2006; **174:** 331–38.

105 Nuermberger EL, Yoshimatsu T, Tyagi S, et al. Moxifloxacin-containing regimens of reduced duration produce a stable cure in murine tuberculosis. *Am J Respir Crit Care Med* 2004; **170:** 1131–34.

106 Rebollo MJ, San Juan GR, Folgueira D, et al. Blood and urine samples as useful sources for the direct detection of tuberculosis by polymerase chain reaction. *Diagn Microbiol Infect Dis* 2006; **56:** 141–46.

107 Torrea G, Van de Perre P, Ouedraogo M, et al. PCR-based detection of the Mycobacterium tuberculosis complex in urine of HIV-infected and uninfected pulmonary and extrapulmonary tuberculosis patients in Burkina Faso. *J Med Microbiol* 2005; **54:** 39–44.

108 Kafwabulula M, Ahmed K, Nagatake T, et al. Evaluation of PCR-based methods for the diagnosis of tuberculosis by identification of mycobacterial DNA in urine samples. *Int J Tuberc Lung Dis* 2002; **6:** 732–37.

109 Aceti A, Zanetti S, Mura MS, et al. Identification of HIV patients with active pulmonary tuberculosis using urine based polymerase chain reaction assay. *Thorax* 1999; **54:** 145–46.

110 Kumar V, Jakubiec W, Li X, et al. Safety, tolerability, pk, and whole blood bactericidal activity (WBA) against *Mycobacterium tuberculosis* of single ascending doses of PNU-100480. *ICAAC* 2009; **49:** F1–1217a (abstr).

111 Williams KN, Stover CK, Zhu T, et al. Promising anti-tuberculosis activity of the oxazolidinone PNU-100480 relative to linezolid in the murine model. *Antimicrob Agents Chemother* 2008; **53:** 1314–19.

112 Kim K, Visintin I, Alvero AB, Mor G. Development and validation of a protein-based signature for the detection of ovarian cancer. *Clin Lab Med* 2009; **29:** 47–55.

113 Chee CB, KhinMar KW, Gan SH, et al. Effect of TB treatment on T-cell interferon-{gamma} responses to M. tb-specific antigens. *Eur Respir J* 2009; published online Nov 19. DOI:10.1183/09031936.00151309.

114 Jafari C, Thijsen S, Sotgiu G, et al. Bronchoalveolar lavage enzyme-linked immunospot for a rapid diagnosis of tuberculosis: a Tuberculosis Network European Trials group study. *Am J Respir Crit Care Med* 2009; **180:** 666–73.

115 Losi M, Bossink A, Codecasa L, et al. Use of a T-cell interferon-gamma release assay for the diagnosis of tuberculous pleurisy. *Eur Respir J* 2007; **30:** 1173–79.

116 Thomas MM, Hinks TS, Raghuraman S, et al. Rapid diagnosis of Mycobacterium tuberculosis meningitis by enumeration of cerebrospinal fluid antigen-specific T-cells. *Int J Tuberc Lung Dis* 2008; **12:** 651–57.

117 Azzurri A, Kanaujia GV, Sow OY, et al. Serological markers of pulmonary tuberculosis and of response to anti-tuberculosis treatment in a patient population in Guinea. *Int J Immunopathol Pharmacol* 2006; **19:** 199–208.

118 Silva VM, Sardella IG, Luiz RR, et al. Immunoreactivity of five antigens of Mycobacterium tuberculosis in patients attending a public health care facility in an area with high endemicity for TB. *Microbiol Immunol* 2008; **52:** 544–550.

119 Geldmacher C, Schuetz A, Ngwenyama N, et al. Early depletion of Mycobacterium tuberculosis-specific T helper 1 cell responses after HIV-1 infection. *J Infect Dis* 2008; **198:** 1590–98.

120 Ewer K, Millington KA, Deeks JJ, Alvarez L, Bryant G, Lalvani A. Dynamic antigen-specific T-cell responses after point-source exposure to Mycobacterium tuberculosis. *Am J Respir Crit Care Med* 2006; **174:** 831–39.

121 Higuchi K, Harada N, Mori T. Interferon-gamma responses after isoniazid chemotherapy for latent tuberculosis. *Respirology* 2008; **13:** 468–72.

122 Chee CB, KhinMar KW, Gan SH, Barkham TM, Pushparani M, Wang YT. Latent tuberculosis infection treatment and T-cell responses to Mycobacterium tuberculosis-specific antigens. *Am J Respir Crit Care Med* 2007; **175:** 282–87.

123 Herrmann JL, Belloy M, Porcher R, et al. Temporal dynamics of interferon gamma responses in children evaluated for tuberculosis. *PLoS One* 2009; **4:** e4130.

124 Pai M, Joshi R, Dogra S, et al. Serial testing of health care workers for tuberculosis using interferon-{gamma} assay. *Am J Respir Crit Care Med* 2006; **174:** 349–55.

125 Pai M, Joshi R, Dogra S, et al. Persistently elevated T cell interferon-g responses after treatment for latent tuberculosis infection among health care workers in India: a preliminary report. *J Occup Med Toxicol* 2006; **1:** 7.

126 Wilkinson KA, Kon OM, Newton SM, et al. Effect of treatment of latent tuberculosis infection on the T cell response to Mycobacterium tuberculosis antigens. *J Infect Dis* 2006; **193:** 354–59.

127 Goletti D, Parracino MP, Butera O, et al. Isoniazid prophylaxis differently modulates T-cell responses to RD1-epitopes in contacts recently exposed to Mycobacterium tuberculosis: a pilot study. *Respir Res* 2007; **8:** 5.

128 Elias D, Akuffo H, Britton S. PPD induced in vitro interferon gamma production is not a reliable correlate of protection against Mycobacterium tuberculosis. *Trans R Soc Trop Med Hyg* 2005; **99:** 363–68.

129 Darrah PA, Patel DT, De Luca PM, et al. Multifunctional TH1 cells define a correlate of vaccine-mediated protection against Leishmania major. *Nat Med* 2007; **13:** 843–50.

130 Hawkridge A, Hatherill M, Little F, et al. Efficacy of percutaneous versus intradermal BCG in the prevention of tuberculosis in South African infants: randomised trial. *BMJ* 2008; **337:** a2052.

131 Floto RA, MacAry PA, Boname JM, et al. Dendritic cell stimulation by mycobacterial Hsp70 is mediated through CCR5. *Science* 2006; **314**: 454–58.

132 Electronic Code of Federal Regulations. Title 21: food and drugs. Subpart H—accelerated approval of new drugs for serious or life-threatening illnesses. March, 2010. http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=783fca020c8214aec9e6fb4cc4871278&rgn=div6&view=text&node=21:5.0.1.1.4.8&idno=21 (accessed March 26, 2010).

133 US Department of Health and Human Services, Food and Drug Administartion. Innovation or stagnation: challenge and opportunity on the critical path to new medical products. Bethesda: Food and Drug Administration, 2004. http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/BudgetReports/2006FDABudgetSummary/UCM148086.pdf (accessed March 26, 2010).

134 Zerhouni E. Medicine. The NIH Roadmap. *Science* 2003; **302**: 63–72.

135 Wagner JA. Overview of biomarkers and surrogate endpoints in drug development. *Dis Markers* 2002; **18**: 41–46.

136 Lathia CD, Amakye D, Dai W, et al. The value, qualification, and regulatory use of surrogate end points in drug development. *Clin Pharmacol Ther* 2009; **86**: 32–43.

137 Lönnroth K, Castro KG, Chakaya JM, et al. Tuberculosis control and elimination 2010–50: cure, care, and social development. *Lancet* 2010; published online May 19. DOI:10.1016/S0140-6736(10)60483-7.

138 Davies PD, Pai M. The diagnosis and misdiagnosis of tuberculosis. *Int J Tuberc Lung Dis* 2008; **12:** 1226–34.

139 Perkins MD, Cunningham J. Facing the crisis: improving the diagnosis of tuberculosis in the HIV era. *J Infect Dis* 2007; **196** (suppl 1): S15–27.

140 Getahun H, Harrington M, O'Brien R, Nunn P. Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings: informing urgent policy changes. *Lancet* 2007; **369:** 2042–49.

141 Reid MJ, Shah NS. Approaches to tuberculosis screening and diagnosis in people with HIV in resource-limited settings. *Lancet Infect Dis* 2009; **9:** 173–84.

142 Perkins MD, Small PM. Partnering for better microbial diagnostics. *Nat Biotechnol* 2006; **24:** 919–21.

143 WHO. Special Programme for Researchand Training in Tropical Diseases (TDR) and Foundation for Innovative New Diagnostics (FIND). Diagnostics for tuberculosis. Global demand and market potential. Geneva: World Health Organization, 2006.

144 Vitoria M, Granich R, Gilks CF, et al. The global fight against HIV/AIDS, tuberculosis, and malaria: current status and future perspectives. *Am J Clin Pathol* 2009; **131:** 844–48.

145 Perkins MD, Bell DR. Working without a blindfold: the critical role of diagnostics in malaria control. *Malar J* 2008; **7** (suppl 1): S5.

146 Girardi E, Sabin CA, Monforte AD. Late diagnosis of HIV infection: epidemiological features, consequences and strategies to encourage earlier testing. *J Acquir Immune Defic Syndr* 2007; **46** (suppl 1): S3–8.

147 WHO. Global tuberculosis control 2009—epidemiology, strategy, financing. Geneva: World Health Organization, 2009.

148 Wallis RS, Doherty TM, Onyebujoh P, et al. Biomarkers for tuberculosis disease activity, cure and relapse. *Lancet Infect Dis* 2009: **9:** 162–72.

149 Young DB, Perkins MD, Duncan K, Barry CE 3rd. Confronting the scientific obstacles to global control of tuberculosis. *J Clin Invest* 2008; **118:** 1255–65.

150 Kaufmann SH. How can immunology contribute to the control of tuberculosis? *Nat Rev Immunol* 2001; **1:** 20–30.

151 Young DB, Gideon HP, Wilkinson RJ. Eliminating latent tuberculosis. *Trends Microbiol* 2009; **17:** 183–88.

152 Barry CE 3rd, Boshoff HI, Dartois V, et al. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol* 2009; **7:** 845–55.

153 Mack U, Migliori GB, Sester M, et al. LTBI: latent tuberculosis infection or lasting immune responses to M. tuberculosis? A TBNET consensus statement. *Eur Respir J* 2009; **33:** 956–73.

154 Andersen P, Doherty TM, Pai M, Weldingh K. The prognosis of latent tuberculosis: can disease be predicted? *Trend Mol Med* 2007; **13:** 175–82.

155 Pai M. Spectrum of latent tuberculosis: existing tests cannot resolve underlying phenotypes. *Nat Rev Microbiol* 2010; **8:** 242.

156 Hill PC, Jackson-Sillah DJ, Fox A, et al. Incidence of tuberculosis and the predictive value of ELISPOT and Mantoux tests in Gambian case contacts. *PLoS One* 2008; **3:** e1379.

157 Bakir M, Millington KA, Soysal A, et al. Prognostic value of a T-cell-based, interferon-gamma biomarker in children with tuberculosis contact. *Ann Intern Med* 2008; **149:** 777–87.

158 Diel R, Loddenkemper R, Meywald-Walter K, Niemann S, Nienhaus A. Predictive value of a whole blood IFN-gamma assay for the development of active tuberculosis disease after recent infection with Mycobacterium tuberculosis. *Am J Respir Crit Care Med* 2008; **177:** 1164–70.

159 Kik SV, Franken WP, Mensen M, et al. Predictive value for progression to tuberculosis by IGRA and TST in immigrant contacts. *Eur Respir J* 2009; published online Oct 19. DOI:10.1183/09031936.00098509.

160 del Corral H, Paris SC, Marin ND, et al. IFNg response to *Mycobacterium tuberculosis*, risk of infection and disease in household contacts of tuberculosis patients in Colombia. *PLoS One* 2009; **4:** e8257.

161 Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. *PLoS Med* 2009; **6:** e2.

162 Pai M, O'Brien R. New diagnostics for latent and active tuberculosis: state of the art and future prospects. *Semin Respir Crit Care Med* 2008; **29:** 560–68.

163 Perkins MD, Roscigno G, Zumla A. Progress towards improved tuberculosis diagnostics for developing countries. *Lancet* 2006; **367:** 942–43.

164 WHO. Rapid tests for drug-resistant TB to be available in developing countries. Geneva: World Health Organization, 2008. http://www.who.int/tb/features_archive/mdrtb_rapid_tests/en/index.html (accessed Feb 3, 2010).

165 Wenner M. New plan seeks to accelerate African diagnostic capacity. *Nat Med* 2009; **15:** 978.

166 WHO and Stop TB Partnership. New laboratory diagnostic tools for tuberculosis control. Geneva: World Health Organization, 2008.

167 Stop TB Partnership and WHO. The Global Plan to Stop TB 2006–2015. Geneva: World Health Organization, 2006.

168 Pai M, Minion J, Sohn H, Zwerling A, Perkins M. Novel and improved technologies for tuberculosis diagnosis: progress and challenges. *Clin Chest Med* 2009; **30:** 701–16.

169 WHO. The use of liquid medium for culture and DST. Geneva: World Health Organization, 2007. http://www.who.int/tb/research/retooling/en/index.html (accessed March 24, 2010).

170 WHO. Policy statement. Molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis (MDR-TB). Geneva: World Health Organization, 2008. http://www.who.int/tb/features_archive/policy_statement.pdf (accessed Feb 3, 2010).

171 WHO. Report of the 9th meeting of the Strategic and Technical Advisory Group on Tuberculosis (STAG-TB). Geneva: World Health Organization, 2009. http://www.who.int/tb/advisory_bodies/stag/en/index.html (accessed March 24, 2010).

172 Pai M, Ramsay A, O'Brien R. Evidence-based tuberculosis diagnosis. *PLoS Med* 2008; **5:** e156.

173 WHO. Reduction of number of smears for the diagnosis of pulmonary TB. Geneva: World Health Organization, 2008. http://www.who.int/tb/dots/laboratory/policy/en/index2.html (accessed Feb 3, 2010).

174 WHO. Definition of a new sputum smear-positive TB case. Geneva: World Health Organization, 2008. http://www.who.int/tb/dots/laboratory/policy/en/index1.html (accessed Feb 3, 2010).

175 Ling DI, Zwerling A, Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *Eur Respir J* 2008; **32:** 1165–74.

176 Steingart KR, Henry M, Ng V, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis* 2006; **6:** 570–81.

177 Minion J, Sohn H, Pai M. Light-emitting diode technologies for TB diagnosis: what's on the market? *Expert Rev Med Devices* 2009; **6:** 341–45.

178 Ramsay A, Cuevas LE, Mundy CJ, et al. New policies, new technologies: modelling the potential for improved smear microscopy services in Malawi. *PLoS One* 2009; **4:** e7760.

179 WHO. Moving research findings into new WHO policies. Geneva: World Health Organization, 2008. http://www.who.int/tb/dots/laboratory/policy/en/index4.html (accessed Feb 3, 2010).

180 Stop TB Partnership's New Diagnostics Working Group and WHO. Pathways to better diagnostics for tuberculosis: a blueprint for the development of TB diagnostics. Geneva: World Health Organization, 2009. http://www.stoptb.org/wg/new_diagnostics/ (accessed March 24, 2010).

181 Pai M, Ramsay A, O'Brien R. Comprehensive new resource for evidence-based TB diagnosis. *Expert Rev Mol Diagn* 2009; **9:** 637–39.

182 Helb D, Jones M, Story E, et al. Rapid detection of *Mycobacterium tuberculosis* and rifampicin-resistance using on-demand, near patient technology. *J Clin Microbiol* 2010; **48:** 229–37.

183 Pai M, Zwerling A, Menzies D. T-cell based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008; **149:** 177–84.

184 Pai M. Guidelines on IGRAs: concordant or discordant? 2nd global symposium on IGRAs; Dubrovnik, Croatia; 2009. http://www.igrasymposium.com/agenda.html (accessed March 24, 2010).

185 Reid A, Scano F, Getahun H, et al. Towards universal access to HIV prevention, treatment, care, and support: the role of tuberculosis/HIV collaboration. *Lancet Infect Dis* 2006; **6:** 483–95.

186 Hopewell PC, Pai M, Maher D, Uplekar M, Raviglione MC. International standards for tuberculosis care. *Lancet Infect Dis* 2006; **6:** 710–25.

187 Yager P, Domingo GJ, Gerdes J. Point-of-care diagnostics for global health. *Ann Rev Biomed Eng* 2008; **10:** 107–44.

188 Usdin M, Guillerm M, Chirac P. Neglected tests for neglected patients. *Nature* 2006; **441:** 283–84.

189 Médecins Sans Frontières. Paris meeting on TB point-of-care test specifications. 2009. http://www.msfaccess.org/TB_POC_Parismeeting/ (accessed March 24, 2010).

190 Marais BJ, Pai M. New approaches and emerging technologies in the diagnosis of childhood tuberculosis. *Paediatr Respir Rev* 2007; **8:** 124–33.

191 Steingart KR, Dendukuri N, Henry M, et al. Performance of purified antigens for serodiagnosis of pulmonary tuberculosis: a meta-analysis. *Clin Vaccine Immunol* 2009; **16:** 260–76.

192 Steingart KR, Henry M, Laal S, et al. Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: a systematic review. *PLoS Med* 2007; **4:** e202.

193 Daley P, Michael JS, Hmar P, et al. Blinded evaluation of commercial urinary lipoarabinomannan for active tuberculosis: a pilot study. *Int J Tuberc Lung Dis* 2009; **13:** 989–95.

194 Mutetwa R, Boehme C, Dimairo M, et al. Diagnostic accuracy of commercial urinary lipoarabinomannan detection in African tuberculosis suspects and patients. *Int J Tuberc Lung Dis* 2009; **13:** 1253–59.

195 Reither K, Saathoff E, Jung J, et al. Low sensitivity of a urine LAM-ELISA in the diagnosis of pulmonary tuberculosis. *BMC Infect Dis* 2009; **9:** 141.

196 Shah M, Variava E, Holmes CB, et al. Diagnostic accuracy of a urine lipoarabinomannan test for tuberculosis in hospitalized patients in a high HIV prevalence setting. *J Acquir Immune Defic Syndr* 2009; **52:** 145–51.

197 Lawn SD, Edwards D, Kranzer K, Vogt M, Bekker LG, Wood R. Urine lipoarabinomannan assay for tuberculosis screening before antiretroviral therapy diagnostic yield and association with immune reconstitution disease. *AIDS* 2009; **23:** 1875–80.

198 Green C, Huggett JF, Talbot EA, Mwaba P, Reither K, Zumla AI. Rapid diagnosis of tuberculosis through the detection of mycobacterial DNA in urine by nucleic acid amplification methods. *Lancet Infect Dis* 2009; **9:** 505–11.

199 Fontela PS, Pai NP, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PLoS One* 2009; **4:** e7753.

200 Sohn H, Minion J, Albert H, Dheda K, Pai M. TB diagnostic tests: how do we figure out their costs? *Exp Rev Anti-infective Ther* 2009; **7:** 723–33.

201 Stop TB Partnership. TB Research Movement. 2008 http://www.stoptb.org/researchmovement (accessed Feb 3, 2010).

202 Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **336:** 924–26.

203 Schunemann HJ, Oxman AD, Brozek J, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. *Evid Based Med* 2008; **13:** 162–63.

204 Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336:** 1106–10.

205 Pai M, Minion J, Steingart KR, Ramsay A. New and improved tuberculosis diagnostics: evidence, policy, practice and impact. *Curr Opin Pulm Med* 2010; **16:** 271–84.

206 Chaisson RE, Harrington M. How research can help control tuberculosis. *Int J Tuberc Lung Dis* 2009; **13:** 558–68.

207 Treatment Action Group & Stop TB Partnership. Tuberculosis Research & Development: 2009 Report on Tuberculosis Research Funding Trends, 2005–2008. New York: Treatment Action Group, 2009.

208 Keeler E, Perkins MD, Small P, et al. Reducing the global burden of tuberculosis: the contribution of improved diagnostics. *Nature* 2006; **444** (suppl 1)**:** 49–57.

209 Dowdy DW, Chaisson RE. The persistence of tuberculosis in the age of DOTS: reassessing the effect of case detection. *Bull World Health Organ* 2009; **87:** 296–304.

210 Dowdy DW, Chaisson RE, Maartens G, Corbett EL, Dorman SE. Impact of enhanced tuberculosis diagnosis in South Africa: a mathematical model of expanded culture and drug susceptibility testing. *Proc Natl Acad Sci USA* 2008; **105:** 11293–98.

211 Dowdy DW, Chaisson RE, Moulton LH, Dorman SE. The potential impact of enhanced diagnostic techniques for tuberculosis driven by HIV: a mathematical model. *AIDS* 2006; **20:** 751–62.

212 Dowdy DW, Lourenco MC, Cavalcante SC, et al. Impact and cost-effectiveness of culture for diagnosis of tuberculosis in HIV-infected Brazilian adults. *PLoS One* 2008; **3:** e4057.

213 Abu-Raddad LJ, Sabatelli L, Achterberg JT, et al. Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *Proc Natl Acad Sci USA* 2009; **106:** 13980–85.

214 Currie CS, Floyd K, Williams BG, Dye C. Cost, affordability and cost-effectiveness of strategies to control tuberculosis in countries with high HIV prevalence. *BMC Public Health* 2005; **5:** 130.

215 Basu S, Friedland GH, Medlock J, et al. Averting epidemics of extensively drug-resistant tuberculosis. *Proc Natl Acad Sci USA* 2009; **106:** 7672–77.

216 Oxlade O, Schwartzman K, Behr M, et al. Global tuberculosis trends: a reflection of changes in tuberculosis control or in population health? *Int J Tuberc Lung Dis* 2009; **13:** 1238–46.

217 Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc Sci Med* 2009; **68**: 2240–46.

218 Yew WW, Chan CK, Leung CC, et al. Comparative roles of levofloxacin and ofloxacin in the treatment of multidrug-resistant tuberculosis: preliminary results of a retrospective study from Hong Kong. *Chest* 2003; **124:** 1476–81.

LETTER FROM INDIA

A DEADLY MISDIAGNOSIS

Is it possible to save the millions of people who die from TB?

BY MICHAEL SPECTER

Every afternoon at about four, a slight woman named Runi slips out of the cramped, airless room that she shares with her husband and their sixteen children. She skirts the drainage ditch in front of the building, then walks toward the pile of hardened dung cakes that people in this slum on the edge of the northeastern Indian city of Patna use for fuel. Dressed in a bright-yellow sari shot with gold threads, Runi is followed by several of her children. Although she can't remember their ages, or her own, Runi must be about forty, because she dates her life from its first crucial memory: the smallpox epidemic that devastated Patna and much of surrounding Bihar province in 1974.

Runi survived that plague, and several others, but, about a year ago, after developing a persistent cough, she visited one of the private medical clinics that line the streets of Patna. There someone who called himself a doctor stuck a needle in her arm, drew a few drops of blood, examined them, and told her that she had tuberculosis. It is not an uncommon diagnosis. Tuberculosis has always been the signature disease of urban poverty, passed easily in poorly ventilated spaces. India has nearly two million new cases each year, and every day a thousand people die of the disease, the highest number in the world. Tuberculosis is also the leading cause of death among people between fifteen and forty-five—the most productive age group in any country and the key to India's prospects for continued economic growth.

For most patients, the choices are bleak. Public hospitals are so overcrowded that people are forced to rely on inaccurate tests dispensed at private labs and clinics. They are unregulated enterprises, and peddle blood tests that are responsible for tens of thousands of misdiagnoses every year. "This is deadly," L. S. Chauhan, the director of the National TB Control Program, told me when we met in New Delhi. "But there are thousands of labs. Shut one down and the next day ten more appear."

Runi's test was indeed worthless. It determined the presence of antibodies, which show that a body's immune system has begun to respond to an infection. But most TB infections are latent: no more than ten per cent will ever cause illness. This means that ninety per cent of people with antibodies for TB in their blood don't have the disease. Runi's cough was clearly caused by something else.

Vaccines and antibiotics have long been seen as touchstones of medical progress. To stop tuberculosis, however, particularly in the developing world, an accurate diagnostic exam is needed even more. In India, China, and Africa, at least two billion people have latent infections. Yet every day thousands are told, mistakenly, that they are sick and need treatment. That's what happened to Runi. Soon after she received her diagnosis, Runi began a regimen of powerful (and toxic) drugs provided by the public-health service, and she stuck to the program for the required six months. Not long after finishing, however, she started to feel worse than she ever had before. "This is the tragedy of our TB-control program," Shamim Mannan said as we watched Runi's children play. Mannan, who is from Assam, a few hundred miles from Patna, serves as the Indian government's chief TB consultant in the region.

"Officially, she is cured," he said. "But how would we know? She took a test that showed she had the antibody for TB in her blood. So do I. So do five hundred million Indians." As Runi stooped to gather fuel for the stove, she began to cough, lightly at first and then with alarming force. Every cough sounded as if somebody had shattered a pane of glass.

"Now she really is sick," he continued, explaining that Runi's TB was no longer dormant, and that taking drugs when they are not necessary often makes them ineffective when they are. "This is what happens when tests mislead us. She will need the drugs again. If they don't work properly, she will be in real trouble. She has almost certainly infected some of her children. That makes everything harder, more expensive, more painful."

Tuberculosis strikes vulnerable people with special ferocity. Victims are seized by severe night sweats, wasted by fatigue, and punished by the blood-tinged cough that is the disease's defining symbol. In most cases, tuberculosis affects the lungs, but it can invade almost any organ of the body. When an infectious person coughs, sneezes, spits, or even shouts, he sends minute particles of sputum, or phlegm, into the air—exposing anyone nearby. For many years, the disease, which is caused by *Mycobacterium tuberculosis*, was referred to as "consumption," because without effective treatment patients often wasted away.

To fight the infection, the body's immune system forms a scar around the TB bacteria which serves as a kind of moat. Afterward, the bacteria lie dormant and cannot spread or infect others. But immune systems fail, and when that happens TB can move from the lungs to the bloodstream and then to the kidneys, the brain, and other organs. (That's why in patients with H.I.V., which ravages the cells that the body uses to defend itself, tuberculosis becomes particularly deadly.) The only way to cure the disease is with a combination of antibiotics. The treatment lasts six months because the drugs work only when the TB bacteria—which grow slowly—are dividing.

For centuries, tuberculosis has been the source of misguided stereotypes, including the association of consumption with creativity and brilliance. "Doctors suspect that tuberculosis develops genius," a 1940 article in *Time* pointed out, "because 1) apprehension of death inspires a burning awareness of life's beauty, significance, transience, 2) the bacillus breeds restlessness and an intoxicated hypersensitiveness." Keats,

Chekhov, the Brontë sisters, and George Orwell—who was born not far from Patna, where his father managed the regional opium trade—all died of the disease.

Nonetheless, tuberculosis has always taken its most serious toll on the industrial-labor class—not on artists. The rise of industry throughout the world has been mirrored uncannily by a rise in deaths from tuberculosis. It was the leading cause of death in Europe and the United States from the eighteenth century into the twentieth. Then prosperity—rather than medicine—drove the rate of infection down. As a society becomes richer, the conditions that allow tuberculosis to flourish start to wane. Sanitation and housing improve and so does nutrition. By the nineteen-fifties, very few people in the West were dying of the disease.

In the developing world, though, tuberculosis has surged dangerously, and this year, according to the World Health Organization, there will be ten million new cases, the largest number in history. As people join the great migrations from villages to crowded cities, slum life and tuberculosis await them. With India's urban population expected to double in the next thirty years, to seven hundred million, its cities will remain fertile ground for an infectious epidemic. Yet—no doubt owing to the fact that rich people in the West rarely get the disease—tuberculosis receives fewer resources, fewer research dollars, and less attention from the global health community than either AIDS or malaria—the two other most deadly infectious diseases. TB activists don't march on Washington or chain themselves to the gates of pharmaceutical firms to demand better treatment.

Tuberculosis can be cured, but taking several antibiotics nearly every day for six months is not easy, particularly in parts of the world without running water or refrigeration. In 1994, the W.H.O. instituted a program called DOTS, which stands for "directly observed treatment, short course." DOTS requires health workers to provide medicine—and then to watch people swallow it every day until they complete their treatment. Compliance is essential, because stopping treatment in the middle permits the most resilient strains of the bacteria to thrive, greatly increasing the chance that they will become resistant to basic, inexpensive drugs.

Thirty-six million people have received care under the DOTS program, eight million of whom would have died without it. It has been a triumph by any measure. Even DOTS, though, has not been able to keep the disease from spreading. That is largely because there is no cheap, reliable test that can determine who is sick and who is not.

Blood tests, like the one Runi had, often do more harm than good. One recent study found that Indians undergo more than 1.5 million useless TB tests of this kind every year. Other approaches are almost as unreliable. Examining a person's sputum—a diagnostic procedure that was developed more than a century ago—remains the most common way to detect the infection. It is a laborious process. Technicians smear the sputum on a slide and then place the specimen under a microscope.

Hospital Road in Darbhanga is home to dozens of unregulated doctors and drug wholesalers. Photograph by Lynsey Addario.

VII NETWORK

The instructions are comically complex. "Spread sputum on the slide using a broomstick," a typical recipe, posted on the wall of a clinic in Patna, begins. "Allow the slide to air dry for fifteen to thirty minutes. Fix the slide by passing it over a flame from three to five times for three to four seconds each time." If the slide isn't held over the flame long enough, false stains will appear—suggesting that people are sick when they are not. Hold the slide too long, though, and the stain will disappear and show nothing at all. The results are accurate little more than half the time.

"You can treat a lot of people, and India has," said Madhukar Pai, an epidemiologist at McGill University and the co-chairman of the international group that assesses new diagnostics for the Stop TB Partnership. "But if you have tests that cause misdiagnosis on a massive scale you are going to have a serious problem. And they do."

Medicine rarely provides magic bullets, but, for the first time, a technology has been developed that might help countries like India escape the endless cycle of mistaken diagnoses and haphazard treatment. A company called Cepheid, based in Sunnyvale, California, now makes a device, called a GeneXpert, that allows doctors to diagnose TB in under two hours—without error or doubt. "The machine is so powerful that it could help end tuberculosis," Mannan told me. "I don't think that is an exaggeration."

An editorial three months ago in the *New England Journal of Medicine* also raised the possibility that, with proper use of this device, tuberculosis—a disease that has been around since the days of the Pharaohs—could be eliminated. The cost, however, would be far too high for the Indian Ministry of Health. "Private business would have to take the lead," Mannan said. "In the past, countries waited until they got richer and tuberculosis mostly went away. India cannot do that. The epidemic is just too big. And we are too poor."

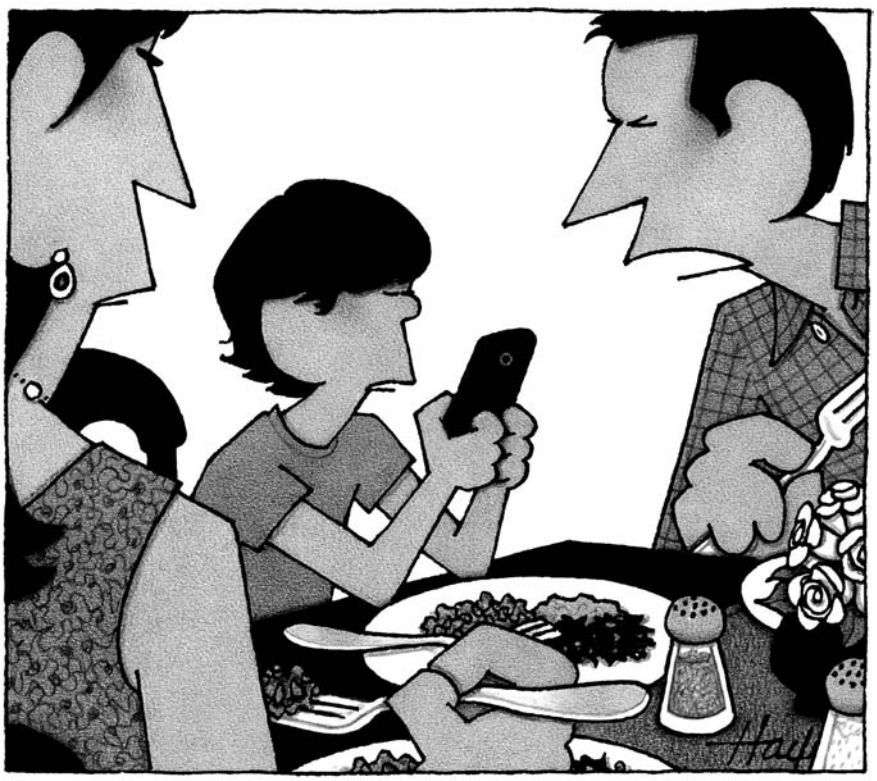
The GeneXpert was developed in 2002, with initial support from the Department of Defense. After the events of September 11th and the mailing of anthrax spores later that year, biological threats became a national priority. The only sure way to recognize dangerous new organisms, whether made by man or by nature, is to analyze their unique DNA, and the GeneXpert has tested billions of pieces of mail for toxins. Its diagnostic capabilities seemed even more promising, however. In 2008, with funding from the Bill and Melinda Gates Foundation, the Foundation for Innovative New Diagnostics, and the National Institutes of Health, researchers at medical centers throughout the world began to assess the machine's effectiveness in diagnosing tuberculosis.

Its success was striking. In a study published along with that editorial in the *Journal*, researchers reported that the GeneXpert identified more than ninety-eight per cent of active TB infections, including many that sputum smears had missed. Because the test looks for the TB bacterium itself, rather than for antibodies, latent infections don't confuse the GeneXpert as they do blood tests. The machine costs nearly twenty-five thousand dollars and each test is about twenty dollars. Prices could plunge if similar machines were introduced and used widely.

"This is absolutely transformational technology," Peter Small, the director of tuberculosis programs for the Gates Foundation, said. "It is a system that removes the guesswork from one of our most deadly diseases." Unlike the sputum technique, the molecular approach is straightforward: a patient spits into a cup, and the sample is placed in a cartridge that looks much like the pods used in many espresso machines. A computer examines the sample's DNA to see if it contains the genetic signature of TB. Results are available within hours.

The GeneXpert can even determine whether the bacteria are resistant to rifampicin, the most effective and widely used component of the four-drug cocktail commonly prescribed for TB. "People often equate sophisticated science with complexity, and this is just the opposite," Small said. "As long as there is electricity, the tests could be carried out by unskilled workers in any village. Training them would be easy, and the potential benefits—saving billions of dollars and millions of lives—worth any effort. The question is how do we get there. I have heard people say that we should trust the government bureaucracy. But others say let's put our faith in an unregulated collection of free agents. It's hard to know which approach is more ludicrous."

I put that question to Mannan, the



"Hey! Elbows off the table."

official responsible for TB control in the Bihar region. A slight, intense man with eyes the color of wet coal, Mannan is a former Army doctor who left the service after he injured his leg jumping from an airplane. He has been frustrated by how rarely the promise of Indian medicine is realized, and by how little entrepreneurs—in one of the world's most entrepreneurial countries—are doing to help.

"We do know that private enterprise can work in India," he said. "Just look at the mobile-phone industry. And the public efforts to halt major diseases have been remarkable. But how do we get them to work together?" Nobody has an answer to that question. The interplay between public and private medicine in India is difficult to navigate, in part because the quality of private medicine varies so wildly. To demonstrate the range of medical options open to most people in Bihar, Mannan suggested that we travel to Darbhanga, about ninety miles northeast of Patna. Before we left, he said, "Everything you find in the country, the good and the bad—it is all in Darbhanga."

Even at first light, the road that leads from Patna, Bihar's capital, to Darbhanga is impossibly crowded. On the ramp of the Mahatma Gandhi Bridge, which passes over the Ganges and leads north toward Nepal, oxen jostle with motorcycles and giant trucks. On the day I made the trip, the traffic was so heavy on the bridge—at more than three and a half miles, it's one of the longest in the world—that it took an hour just to reach the lush banana plantations on the other side.

Patna and Darbhanga were once important centers of civilization. Buddha found enlightenment under a bodhi tree in Bihar, twenty-five hundred years ago, and the Fortress of Maharajas still stands in Darbhanga. Today, though, the province lags behind other regions of India in every category of economic and human development. Its eighty-five million residents earn, on average, less than half what people in the rest of the country earn; plumbing and sanitation facilities are meagre. Tens of thousands of migrants pass through Darbhanga each year as they abandon their ancestral villages and seek new lives in Delhi, Mumbai, and other major cities.

The Medical College Hospital, an imposing white fortress spread over several city blocks, is the largest in the region, but the city is also home to what may well be India's most formidable collection of unregulated pharmaceutical wholesalers, a kind of medical red-light district. Virtually any drug can be purchased, in whatever quantity one desires, without a prescription. Want a thousand polio vaccines? Narcotic painkillers, cancer medication? Scarce AIDS therapies? They are all readily available in Darbhanga. But rarely at the hospital.

The tuberculosis and AIDS clinics at the Medical College Hospital are open every day from 8 A.M. to 2 P.M. By the time Mannan and I walked into the cavernous waiting room early that morning, patients packed the benches and sprawled across the floor. Most sat silently, their eyes hollow, their heads down. The sound of harsh coughing filled the air. The line for medications snaked into the courtyard, where dozens of women, many of them cradling infants in their arms, waited patiently.

Like other public hospitals in the developing world, the Medical College Hospital struggles to provide medicine for its patients. The dispensary is rudimentary: basic tuberculosis drugs are available, but not those needed to treat resistant strains, which now account for nearly twenty per cent of India's growing caseload. For people who do not respond to the first line of TB treatments, there are two choices: find money to buy medicine somewhere else or get sicker.

Since late 2009, the hospital has had one unique asset: a piece of equipment called a P.C.R., which can multiply tiny samples of DNA and analyze them. The device is not as fast as the GeneXpert, but it can examine the genetics of virtually any organism, including tuberculosis. The hospital's machine, which was purchased with money from a government research grant, has never been used. "The hospital has had this for months," Mannan said. "But nobody knows how it works." We were standing at the door of the virology lab, where the new P.C.R. Cobas TaqMan 48, made by Roche and sold for roughly fifty thousand dollars, was resting on a shelf, still wrapped in its shipping material.

How could that be? I was staring at a machine that could alter, even save, the lives of scores of the people who were sitting nearby in the gathering heat. Mannan said nothing, though his anger was palpable. "Ask them," he said, referring to the scientists who worked in the hospital, when I tried to get him to explain. "They will tell you."

We walked down the hall to meet Ravindra Prasad, a doctor in the department of social medicine. He was an agreeable man with a round face and an easy manner. I asked why the P.C.R. machine sat imprisoned and unused.

"The chemical kit expired," he said, smiling politely. The chemicals used in the machine have a short shelf life; but I learned later that they are not hard to replace. That couldn't have been the reason. "The methods we have for diagnosing tuberculosis all function smoothly," Prasad added, as if he were reading from a prepared statement. He was referring to sputum tests, which are often inaccurate. "We follow the standard manual." Prasad offered us tea, but said nothing more about the medical needs of his patients. "It's a nice lab," Mannan said when we left. "Beautiful, actually. But if the doctors used it properly that would interfere with their private practice."

I asked what he meant.

"It is simple," he said. "If patients are treated at the hospital, they won't need to pay for anything else."

The Darbhanga medical red-light district lies just a few blocks from the main hospital. On most days, as the public clinics prepare to take their last patients, touts appear in the waiting rooms and on the hospital grounds, eager to steer people toward a private doctor on Hospital Road. More than eighty per cent of medical services in India are in private hands, and healthcare costs are among the most common reasons for bankruptcy.

The touts—equal parts salesmen, psychologists, and pimps—are good at their job. If you need TB medication or a test or an X-ray, these men will get you quickly to a clinic that charges for services people are entitled to receive at no cost in public hospitals. According to Mannan, the tout receives ten per cent of any eventual fee from a referral. Rickshaw drivers get five per cent, medical assistants ten, and the referring doctor, almost always a physician based at the Med-

ical College Hospital, thirty-five per cent. That leaves forty per cent for the clinician.

Much of the time, the referring physician from the public hospital is also the private clinician who does the work. That earns him seventy-five per cent of any fee. Public salaries are not sufficient to support most doctors, so, every afternoon, many of the hospital's physicians work in these private clinics.

Well-trained doctors are not the only people working on Hospital Road, however. Officially, a doctor needs a license to practice medicine in India. In fact, though, there are no mechanisms to verify the validity of licenses or to punish people who break the law. It is not rare for "doctors" to lack medical training completely.

We arrived as darkness began to fall; hundreds of people, having finished the workday, crowded the rutted streets. There were dozens of drug shops, with names like Raj Medical Agency, Krishna Scientific and Surgical Works, and Zar Whole Sale Drugs—often illuminated by a single bulb. The streets of the medical red-light district are filled with "specialists." Mannan and I wandered into a back alley where two men asked after our health with more solicitousness than was necessary. I asked what they were offering, and one of them let out a loud cackle.

"Let me show you," he said, and led us to a small room with several chairs, a table, and three refrigerators. The man said that his name was Pranay, and he offered a variety of blood tests, for liver function, kidney function, H.I.V., and several other standard diagnostics, all at reasonable prices. Wholesalers make their money through volume sales, not high prices. "We get twenty-five to thirty referrals a day," he told me.

The stall next door could have been an exhibit in a science museum: it contained an ancient X-ray machine, held together with duct tape and baling wire. The owner had just finished taking chest slides for a middle-aged man. He didn't offer any of the customary lead shields or other protections against possible radiation leaks—and that machine certainly leaked. "It's safe," the man said. "They are X-rays."

He told us that he ran about fifteen to twenty chest X-rays a day; he charges a hundred rupees for each, or a little more than two dollars. His services were also available for broken bones and other routine problems. I asked how he had acquired his equipment and where he had learned to use it. He told us that he had taken the X-ray machine from a hospital in Bihar that was about to throw it away. The idea of training made him laugh. "Did you see 'Slumdog Millionaire'?" he said. "Before this, I was a chai wallah"—a man who serves tea—"just like that kid."

It was time to return to Patna; driving late at night on the roads of rural India is a risky business. Before we left, though, Mannan insisted that we make one more stop, at another clinic nearby. The place was essentially an open concrete garage; against one wall stood a small table with hot plates on which patients could heat rice. The room was full, and more than a dozen people stood on the street, waiting to get in. "This is the best TB clinic in town," a pharmacist who owned the shop next door explained.

The head of the clinic, Dr. P. M. Srivastav, works at Medical College Hospital, and we had spoken with him earlier. At night, for a hundred and thirty rupees, Srivastav will see anyone who waits in line. He doesn't test for tuberculosis at his clinic, and said that he refers people he suspects of having the disease to the hospital. He does, however, earn a fee from every patient he sees, including those he sends back to the hospital for free treatment. "Now do you understand why that machine is wrapped in plastic?" Mannan asked.

As we were about to leave, a large car pulled up at the front door. Srivastav climbed out of the back seat, looked at us with surprise, and smiled sheepishly. Before I had a chance to ask a question, he was gone, safely tucked away in his private office.

The uncertainties and dangers of diagnosis remain the greatest obstacle to successful TB treatment, in India and throughout the developing world. For that to change, investments from international aid organizations and from private companies will be necessary. That may seem unlikely, but it has happened before, most notably with AIDS drugs. In the nineteen-eighties, when AZT became the first effective treatment for H.I.V., the annual cost for each patient was ten thousand dollars. People in the West, who were rich or lucky enough to have good insurance, could afford it. In countries that struggle to provide basic immunizations against diseases like measles, though, AIDS treatments were a fantasy. Then various groups, including the Clinton Foundation, the Gates Foundation, and the Global Fund to Fight AIDS, Tuberculosis and Malaria, joined together to push for lower prices. Generic manufacturers, led by Cipla, the Mumbai-based pharmaceutical giant, began to churn out highly effective medicine at a small fraction of what it cost in the United States. Political pressure mounted, officials of the World Health Organization joined the call for cheaper AIDS medications, and today the governments of poor countries like India can buy those drugs for an annual price of less than a hundred dollars per patient. These drugs are normally distributed in bulk, through international AIDS organizations.

A similar effort will be required to lower the cost of diagnosing tuberculosis. There will also have to be a transformation in how TB medicine is regulated. That may seem like an insurmountable barrier, but, with the proper incentives, the system could work. Again, one can look to the history with AIDS medicines for a model. Because Cipla and other Indian pharmaceutical companies are frequently inspected by international regulators—such as the U.S. Food and Drug Administration—governments are willing to buy their products. That's one reason that Indian firms have become the most important manufacturers of generic AIDS medicines in the world.

Any company that sells molecular diagnostics would need the same sort of oversight. But producing cheap, internationally acceptable versions of the GeneXpert would surely lead to great profits.

"You have to keep in mind that India has many terrible doctors," Mad-

hukar Pai told me. "But it also has some of the best private medicine available." I saw that in Darbhanga, where, in addition to the shoddy purveyors of the medical red-light district, I visited the Geeta Molecular Diagnostic Lab, a new private facility not far from the center of town. There I was greeted by a team of researchers, all in starched lab coats, including Deepak K. Prasad, a geneticist and the director of the laboratory. He led us on a tour: there were separate sections for gene detection, gene amplification, and histological analysis. Geeta Diagnostics had two P.C.R. machines and other, similarly advanced diagnostic tools. Few facilities in New York are better equipped. Patients sat on cream-colored couches reading magazines and sipping tea.

"The genetic approach to diagnosis is really where medicine is going," Prasad told us. The company, which is two years old, offers tests for heart disease, several types of cancer, thyroid disease, H.I.V., and tuberculosis, among other disorders. The TB test costs fifteen hundred rupees—a little more than thirty dollars. The lab does between fifty and seventy-five each week, and its doctors are paid well enough so that they don't need to work at second jobs.

"You can call it expensive, but you have to look at the eventual costs, not the initial price of a single test or one piece of machinery," Prasad said. That would be difficult to dispute. Thirty dollars may be a lot of money for most Indians; but treating drug-resistant strains of tuberculosis costs thousands of dollars and places a terrible burden on the country, not to mention on the people who are sick. In fact, treatment and deaths caused by TB in India cost more than three billion dollars in lost productivity each year.

The power of machines like the GeneXpert has already become evident at Mumbai's Hinduja Hospital, a private institution that has been using one for three years. Mumbai has one of the worst TB problems in India, particularly with drug-resistant cases. Yet at Hinduja the machine has made it possible for doctors to diagnose and treat patients before they are able to spread the disease. "There has always been a pretty standard approach to using fancy medical technology," Camilla Rodrigues, who runs the microbiology department, told me when I visited. "You develop it in the West and use it there. Eventually, it trickles down to the poor countries." Rodrigues pointed out that, with tuberculosis, the pattern makes no sense. The GeneXpert was invented in the West, but India and Africa need it much more urgently. "Every time we make a correct diagnosis, we save not one life but many," she said, waving in the direction of the boxy metal-and-Plexiglas machine sitting in a corner of the lab. "And with this machine we make correct diagnoses in two hours."

Rodrigues has been working with tuberculosis for two decades. "When I started, it seemed hopeless," she said as we sat in her office, which is adjacent to a busy lab filled with graduate students, most of whom are focussing on TB.

"You would ask people why we are not doing more to stop this terrible, crippling epidemic, and the answer was usually a shrug," she continued. Rodrigues has cavernous eyes and long dark hair pulled back in a bun. She speaks frankly but somehow conveys a buoyant sense of optimism. "For so long tuberculosis has been a part of life here. In the past, if you said you have the disease people would hardly flinch. Can you imagine going to a neighbor in New York and saying you have tuberculosis? People would shriek."

Lately, though, Rodrigues has begun to sense a shift away from the habitual fatalism that has defined the Indian approach to public health. "Sometimes I go to Churchgate Station," she continued. "It is the busiest train station in the city, maybe in the country. I go at rush hour. You cannot move or breathe or think. You cannot walk or talk. It is the perfect place to spread tuberculosis.

"But it is also the perfect place to stop it," she said. "I walk around that platform and I look at people and I say to myself, Which of you are sick? We need to know. And, finally, after more than a century we can know. At this point, it is just a matter of will." ♦

NEWYORKER.COM

Michael Specter takes readers' questions, and narrates a slide show of photos from Bihar.

EDITORIALS

Tuberculosis Diagnosis — Time for a Game Change

Peter M. Small, M.D., and Madhukar Pai, M.D., Ph.D.

The effective treatment of tuberculosis is a life-saving intervention. The global scale-up of tuberculosis therapy has averted 6 million deaths over the past 15 years, making it one of the greatest public health interventions of our lifetime.[1] Unfortunately, by the time most patients are treated, they have already infected many others.[2] This failure to interrupt transmission fuels the global epidemic so that every year there are more new cases of tuberculosis than in the previous year.[1]

National tuberculosis programs are particularly challenged by multidrug-resistant tuberculosis. Globally, fewer than 2% of the estimated cases of multidrug-resistant disease are reported to the World Health Organization (WHO) and managed according to international guidelines. The vast majority of the remaining cases are probably never properly diagnosed or treated, further propagating the epidemic of multidrug-resistant tuberculosis. The situation is further worsened by the epidemic of human immunodeficiency virus (HIV), especially in Africa.

For decades there has been little effort to improve techniques for diagnosing tuberculosis.[3,4] Consequently, tuberculosis tests are antiquated and inadequate. The most widely used test (smear microscopy) is 125 years old and routinely misses half of all cases. These inadequacies are particularly problematic since such tests are generally performed in underfunded and dysfunctional health care systems.[4,5] The problem is exacerbated by the widespread use of inaccurate and inappropriate diagnostic tools, such as serologic assays, in many countries.[6]

Fortunately, in the past few years, several improved tuberculosis tests have received WHO endorsement for widespread use.[6,7] In this issue of the *Journal,* Boehme and colleagues[8] describe a new automated nucleic acid–amplification test that may allow a relatively unskilled health care worker to diagnose tuberculosis and detect resistance to a key antibiotic within 90 minutes. This test and others that are likely to follow have the potential to revolutionize the diagnosis of tuberculosis. Thus, in the coming years, rapid diagnosis and targeted treatment will provide the greatest opportunity for stopping the tuberculosis epidemic.

In a large, well-conducted, multicountry study, Boehme et al. evaluated an automated tuberculosis assay (Xpert MTB/RIF) for the presence of *Mycobacterium tuberculosis* (MTB) and resistance to rifampin (RIF). With a single test, this assay identified 98% of patients with smear-positive and culture-positive tuberculosis (including more than 70% of patients with smear-negative and culture-positive disease) and correctly identified 98% of bacteria that were resistant to rifampin.[8]

The assay has several critical advantages over conventional nucleic acid–amplification tests, which have been licensed for nearly 20 years and yet have not had a substantial effect on tuberculosis control. The MTB/RIF assay is simple to perform with minimal training, is not prone to cross-contamination, requires minimal biosafety facilities, and has a high sensitivity in smear-negative tuberculosis (the last factor being particularly relevant in patients with HIV infection).[8]

However promising these findings, issues involving the MTB/RIF assay may limit its global utility. These issues include its high cost, limitations in testing only for rifampin resistance, a platform that detects a relatively small number of mutations, and inability to indicate which patients are "sputum smear–positive" for reporting purposes, infection-control intervention, and treatment monitoring.

On the plus side, the MTB/RIF assay promises to decentralize molecular diagnosis, since it potentially can be used at the point of treatment in a microscopy center or in a tuberculosis or HIV clinic. However, because Boehme et al. used the test at reference laboratories, their study offers only indirect proof of concept for use in such settings. Critical to a rapid scale-up of the test will be the results of additional studies to determine how it performs in such settings and whether its use improves outcomes for patients in a cost-effective manner.

If an improved rapid nucleic acid–amplification test is adopted globally, it could help avert more than 15 million tuberculosis-related deaths by 2050.[9] However, even the most promising diagnostic test will have only limited impact if it does not reach the patients who need it. As with any diagnostic test or intervention, its actual impact will depend on the system in which it is used. Health systems must be strengthened so that patients do not delay in seeking care and have prompt access to appropriate treatment once they receive a diagnosis. Health-system barriers to the use of improved technologies must be anticipated and addressed. Although the burden on health systems will be reduced by a simple dipsticklike, point-of-care assay, such tests are not likely to be available in the short term.[7]

To realize the potential of improved technologies, a diverse set of stakeholders need to support large-scale innovation and delivery. Scientists and industry need to develop radically improved tools, including drugs and vaccines, while offering reasonable pricing that reflects public health needs and economic realities in resource-limited countries. Operational and implementation researchers need to quickly identify and respond to the full spectrum of issues that form the critical path to improving the prevention and control of tuberculosis. Policymakers and regulators must turn scientific evidence into permissive policies and regulations that allow national programs to rapidly incorporate new tools. Funders must increase and reprogram resources to become conduits for innovation and not fund decades-old technologies for years into the future. Programs must maintain focus on the basics of tuberculosis control while quickly modifying delivery systems to take advantage of the benefits of improved tools. Lastly, patient advocates and activists should hold everyone accountable and ensure that communities drive demand for improved systems and tools.

Despite these challenges, it is clear that improvements in diagnostics are driving a virtuous cycle in care: the promise of improved tests drives their uptake, their uptake results in better health outcomes, improved outcomes attract more funding for health care systems, and better-funded systems are an incentive to the development of even better technologies. We are particularly optimistic about the potential role of governments, product developers, and companies in emerging economies with high tuberculosis burdens, such as China, India, Brazil, and South Africa. These countries now have the capacity to develop low-cost generic or novel assays adapted to local contexts and incorporate their scale-up in both national tuberculosis-control programs and private laboratories, supported by successful public–private partnerships. Emerging economies have the potential to become global leaders in innovative product development and delivery. If these countries successfully tackle their own tuberculosis problems, the elimination of tuberculosis by 2050 might become a reality.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

From the Global Health Program, Bill and Melinda Gates Foundation, and the Institute for Systems Biology, Seattle (P.M.S.); and the Department of Epidemiology and Biostatistics, McGill University, and Montreal Chest Institute, Montreal (M.P.).

This article (10.1056/NEJMe1008496) was published on September 1, 2010, at NEJM.org.

1. Lonnroth K, Castro KG, Chakaya JM, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet 2010;375:1814-29.
2. Dye C, Williams BG. The population dynamics and control of tuberculosis. Science 2010;328:856-61.
3. Perkins MD, Small PM. Partnering for better microbial diagnostics. Nat Biotechnol 2006;24:919-21.
4. Small PM. Strengthening laboratory services for today and tomorrow. Int J Tuberc Lung Dis 2008;12:1105-9.
5. Perkins MD, Cunningham J. Facing the crisis: improving the diagnosis of tuberculosis in the HIV era. J Infect Dis 2007; 196:Suppl 1:S15-S27.
6. Pai M, Minion J, Steingart K, Ramsay A. New and improved tuberculosis diagnostics: evidence, policy, practice, and impact. Curr Opin Pulm Med 2010;16:271-84.
7. Wallis RS, Pai M, Menzies D, et al. Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. Lancet 2010;375:1920-37.
8. Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 2010;363:1005-15.
9. Abu-Raddad LJ, Sabatelli L, Achterberg JT, et al. Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. Proc Natl Acad Sci U S A 2009;106:13980-5.

Proposals for a Phased Evaluation of Medical Tests

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

Background. *In drug development, a 4-phase hierarchical model for the clinical evaluation of new pharmaceuticals is well known. Several comparable phased evaluation schemes have been proposed for medical tests.* ***Purpose.*** *To perform a systematic search of the literature, a synthesis, and a critical review of phased evaluation schemes for medical tests.* ***Data Sources.*** *Literature databases of Medline, Web of Science, and Embase.* ***Study Selection and Data Extraction.*** *Two authors separately evaluated potentially eligible papers and independently extracted data.* ***Results.*** *We identified 19 schemes, published between 1978 and 2007. Despite their variability, these models show substantial similarity. Common phases are evaluations of technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal aspects.* ***Conclusions.*** *The evaluation frameworks can be useful to distinguish between study types, but they cannot be seen as a necessary sequence of evaluations. The evaluation of tests is most likely not a linear but a cyclic and repetitive process.* ***Key words:*** *medical tests; biomarkers; test evaluation; medical technology assessment.* ***(Med Decis Making. 2009;29:E13–E21)***

Over the last few decades, many new medical tests have been developed, and the number of available options is still increasing. Premature dissemination of testing technologies can lead to erroneous diagnoses and preventable delays in starting appropriate treatment or, alternatively, to the initiation of unwarranted, sometimes dangerous therapy. Examples have been the dexamethason suppression test for depression, the carcinoembryonic antigen for colon cancer, and the 125I-fibrinogen leg scan for the diagnosis of deep venous thrombosis.[1,2] In addition, the increasing costs of health care have put pressure on available budgets, calling for the elimination of ineffective medical technology. These are ample reasons why new medical tests should be thoroughly evaluated before they are introduced in clinical practice.

The ultimate benefit of any medical technology should be expressed in terms of its effects on health outcome, and tests are no exception.[3] Yet the evaluation of technology can be a time-consuming and costly process. An efficient use of resources calls for a well-planned evaluation strategy. In such a strategy, more elaborate and therefore more expensive forms of evaluation should only be performed if sufficient evidence has been obtained in previous steps of the evaluation process. Such a phased approach, moving gradually

Received 8 September 2008 from the Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (JGL, ML, PMMB), and the Department of Psychiatry, Waterland Hospital, Purmerend, the Netherlands (JGL). Revision accepted for publication 22 February 2009.

Address correspondence to Patrick M. M. Bossuyt, PhD, Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, Room J1b-212, PO Box 22700, 1100 DE Amsterdam, the Netherlands; e-mail: p.m.bossuyt@amc.nl.

DOI: 10.1177/0272989X09336144

This article is part of the White Paper series from the Agency for Health Care Research and Quality (AHRQ) Effective Health Care Program. An earlier version of this article was presented at the Diagnostic Test Evaluation Working Meeting, Rockville, MD, May 28–29, 2008.

from small to larger studies, may also protect the rights and integrity of human volunteers and patients.

In drug development, a 4- or 5-phase hierarchical model for the clinical evaluation of new products is well known. Phase 0 studies are exploratory first-in-human trials to evaluate whether the drug or agent behaves in human subjects as was expected from preclinical studies. In phase I, the safety, tolerability and toxicity, and pharmacodynamics and pharmacokinetics of the new drug are assessed. Phase II usually consists of small-scale clinical investigations to obtain an initial estimate of the effect of treatment. If the treatment effect is too small, further evaluation will be discontinued. In phase III, the effectiveness of the drug is assessed by measuring patient outcome in randomized clinical trials. If the drug is effective, further surveillance after introduction to the market is necessary. In phase IV, the long-term effects and side effects can be registered.

Several comparable hierarchical models have been proposed for the evaluation of diagnostic tests. Analogous to the 4-phase model for the evaluation of new drugs, these models require that in each phase certain conditions be fulfilled before the evaluation can continue with the subsequent phase. Several of these proposals are closely related to hierarchies of evidence. One of the best known are the levels of efficacy for imaging tests, proposed by Fryback and Thornbury in this journal in 1991.[4]

Several more proposals have appeared since then. It is unclear to what extent these models differ and, if so, in what elements. We have performed a systematic search of the literature for phased or hierarchical models for the evaluation of medical tests. We present our findings, a synthesis of existing models, and a critical commentary.

LITERATURE SEARCH

Eligible for this review were papers that described a proposal for the phased evaluation of medical tests, from the first technological laboratory-based evaluation to the evaluation of the performance of the test in clinical practice. Studies that only described parts of this process and studies that advocated a less linear approach were also included in our review.

Papers describing hierarchical models for the evaluation of diagnostic tests use different words and descriptions for these models in their titles and abstracts. In general, these studies are not indexed in a consistent way in electronic bibliographic databases. We first searched in Medline, Web of Science, and Embase for studies with the following words: (phased approach[tiab] OR hierarchical model[tiab] OR phased evaluation[tiab] OR hierarchical approach[tiab] OR hierarchical evaluation-[tiab]) AND (diagnosis[tw] OR diagnostic[tw] OR diagnosis[MeSH]) (239 hits, January 2009).

The databases mentioned were then searched for similar or related articles and for articles that cited the included papers. We also manually checked the reference lists of identified papers. When a paper only made a reference to a previous proposal for a hierarchical evaluation, without further modification, it was excluded.

Models for the Phased Evaluation of Tests

We identified 31 papers with a model for the phased evaluation of diagnostic tests. Two of these were based on a model previously proposed by Guyatt and others. Two others referred to a model of Fineberg, 1 was based on the model of Sackett and Haynes, and 7 papers referred to Fryback and Thornbury. In total, 19 different models were found. The first one of these was published in 1978; the most recent paper appeared in 2007.

The oldest references we could identify appeared in a special issue of the *American Journal of Roentgenology* on the evaluation of computed tomography. When computed tomography was widely adopted in the United States in the early 1970s, it became the focus of much debate on the evaluation of diagnostic imaging and new health technologies in general. In an editorial, Fineberg noted that ''one of the difficulties in evaluating a diagnostic test is its remoteness from health outcome.'' Yet ''the ultimate value of the diagnostic test is that difference in health outcome resulting from the test: In what ways, to what extent, with what frequency, in which patients is health outcome improved because of this test?''[5] Loop and Lusted reported how the American College of Radiology (ACR) had tried to deal with the problems of evaluating the health consequences of testing. The ACR had established an Efficacy Studies Committee in 1972, chaired by Lee B. Lusted. That committee decided that ''the fullest and most long-range expression of efficacy ought to include some measure of the influence of the examination on the final outcome of the episode of ill health.''[6] The committee distinguished between diagnostic efficacy (E-1), the change in the probability of diagnosis after radiographic results have become available, therapeutic efficacy (E-2), the change in therapy planning, and outcome efficacy

(E-3): was the patient better off as a result of the procedure having been performed?

Building on this model, Fryback and Thornbury developed their framework, which appeared in 1991 in a Lusted memorial issue of *Medical Decision Making*.[4] Both authors have described the framework in more detail in later publications.[7,8] Theirs is a 6-tiered hierarchical model, which extends from the physics of imaging, through clinical use in decisions about diagnosis and treatment, to patient outcome and societal issues. Demonstration of efficacy at each lower level in this hierarchy, they wrote, is logically necessary but not sufficient to assure efficacy at higher levels. Kent and Larson used almost the same levels in discussing the efficacy of magnetic resonance imaging but added 2 other dimensions: the spectrum of diseases and the quality of research.[9] Another modification of the ACR framework was proposed by Mackenzie and Dixon.[10] Phelps and Mushlin combined medical decision theory and epidemiological information in suggesting 2 hurdles for diagnostic technologies, linking the accuracy level with the societal level.[11]

Silverstein and others translated the ACR approach to laboratory medicine, and Pearl applied it to tests in general.[12,13] The related ACCE framework for the evaluation of genetic tests is a model process for evaluating data on emerging genetic tests. The acronym is taken from the 4 components: analytical validity; clinical validity; clinical utility; and ethical, legal, and social implications.[14]

Several others have translated the ACR levels of efficacy into phases of evaluation. In 1978, Freedman classified designs to evaluate and compare imaging techniques and observed a parallel with the standard classification of clinical trials.[15] Studies of diagnostic accuracy, he wrote, are analogous to phase II trials, whereas studies evaluating the contribution to clinical management correspond to the phase III category. The majority of studies he observed at the time were phase II type accuracy studies, and more emphasis on phase III studies was required. In a similar way, Taylor and others classified 200 studies published in the *American Journal of Roentgenology* and in *Radiology* in 1988 and 1989 into 1 of 5 phases.[16] They found that the majority of studies focused on early technical assessment.

Guyatt and others from McMaster University also extended the ACR framework into a proposal for stepwise clinical evaluation of diagnostic technologies.[17] Diagnostic technology assessment should begin by establishing the capability of the technology under ideal or laboratory conditions, followed by an exploration of the range of possible uses and the accuracy of the test. Their proposal also contains a very strong plea for randomized clinical trials of test strategies and a critical discussion of some of the poorer study designs. van der Schouw, Van den Bruel, and their respective colleagues similarly suggested stepwise evaluations of tests.[18,19]

Kobberling and others proposed a 4-phased model for test evaluation, explicitly emphasizing the similarity with the evaluation of therapeutic methods.[20] In 2000, Houn and others from the Food and Drug Administration (FDA) noticed a similarity in the evaluation of breast imaging technology and the phased approach of the agency in the clinical development of drugs and biological products.[21] Phase I refers to the initial evaluation of a developing technology in human populations. Phase II refers to clinical studies involving limited numbers of human subjects to gather preliminary evidence regarding effectiveness and additional safety data. Phase III refers to controlled clinical studies intended to provide a reasonable assurance of safety and effectiveness in defined populations. Finally, phase IV refers to studies performed once a technology has gained marketing approval; these studies address long-term safety and better characterize the performance of the technology within a larger population. In an accompanying editorial, Gatsonis introduced a paradigmatic matrix for the evaluation of imaging technology, with 4 phases and 3 possible end points for studies.[22] The 4 phases correspond to what he called the developmental age of the modality, starting from discovery, and then moving to introduction, maturity, and dissemination. In the early phases, the focus is on diagnostic performance, whereas later phases would focus on impact on the process of care and patient outcome.

While schemes inspired by the proposals by Lusted, Fineberg, and Guyatt made a distinction between accuracy, diagnostic impact, and therapeutic impact, other authors have proposed multiphase models for the evaluation of accuracy in itself. Zweig and Robertson suggested the label ''Phase I Trial'' for studies of the analytical precision, accuracy, sensitivity, and specificity of a laboratory test, while ''Phase II Trials'' would refer to studies determining the usual range of results encountered in healthy subjects or comparing the results obtained in various disease states with this usual range.[23] A prospective diagnostic trial of the actual clinical usefulness of a test in a realistic clinical setting would then be termed a ''Phase III Trial.'' Multiple phases in the evaluation of accuracy have also been proposed by Sackett and Haynes,[24] Pepe,[25] and Taube, Jacobson, and Lively.[26] Elsewhere, Obuchowski

Table 1 Summary of Proposals for the Phased Evaluation of Medical Tests

	Loop 1978[6]	Zweig 1982[23]	Guyatt 1986[17]	Freedman 1987[15]	Memorandum 1990[20]	Fryback 1991[4]	Kent 1992[9]	Taylor 1993[16]	Silverstein 1994[12]	van der Schouw 1995[18]	Mackenzie 1995[10]	Pearl 1999[13]	Houn 2000[21]	Gatsonis 2000[22]	Sackett 2002[24]	Haddow 2003[14]	Pepe 2005[25]	Taube 2005[26]	Van den Bruel 2007[19]
Levels/Phases																			
Technical efficacy		1	1	1	1	1	1	1–3		1	1	1	1	1		1	1–3	1–2	1
Intended use			2															3	2
Diagnostic accuracy			3	2	2	2	2	4	1		2	2	2			2	4	4–6	3
Usual range Subgroups		2			3					2				1	1–2				
Clinical population		3			4					3				2	3			7	
Diagnostic thinking efficacy	1		4			3	3		2	4	3	3							
Therapeutic efficacy	2		5			4	4		3		4	4	3	3–4					
Patient outcome efficacy	3		6	3	5	5	5	5	4		5	5	4	3–4	4	3	5		4
Societal efficacy				4		6				5		6	5			4			5

discussed how the questions and the number of readers should vary with a phased evaluation of imaging.[27]

HIERARCHICAL MODELS: A SYNTHESIS

In Table 1, we have summarized the levels and phases described by the 19 different models. Each model consists of 4 to 7 different elements, with marked similarities between these proposals. Most models start with a phase I, which consists of test development. During this phase, the test has to meet prespecified technical requirements. Aspects that have to be documented in this phase include feasibility, required equipment and personnel, and physical and biochemical parameters specific to the test, such as the minimal detection level, circadian fluctuation, resolution, contrast level, and reproducibility. Guyatt and others recommended that, in addition, the test should be applied to a large number of diverse conditions in order to delineate its possible uses.[17]

In most models, the diagnostic accuracy of the test is assessed in one or more subsequent phases. The results of the test under evaluation are compared to those from a reference standard in order to establish how well the test is able to identify patients with the target condition. Diagnostic accuracy can then be characterized in terms of sensitivity and specificity, predictive values, likelihood ratios, or receiver operating characteristic curves and derived measures.

Some authors distinguish a series of subphases at this phase. They propose to evaluate the diagnostic accuracy first in a group of subjects with the disease of interest and a group of healthy persons for an easy comparison. Subsequently, the evaluation is extended to other parts of the disease spectrum. Finally, diagnostic accuracy is evaluated in a clinical study group that closely resembles the population of patients for which the test is intended. In addition, some authors suggest comparing the diagnostic accuracy of the test with the performance of other tests intended to detect the same target condition before proceeding further.

Most proposals continue with the evaluation of the clinical effectiveness of the test, assessed in terms of its effect on diagnostic thinking and patient management, therapeutic efficacy, and patient outcome. To investigate diagnostic thinking efficacy, Fryback and Thornbury suggested studies to document the percentage of cases in which an image was judged ''helpful'' to making the diagnosis or to summarize the difference in clinicians' subjectively estimated diagnosis probabilities before and after receipt of test information.[4]

Studies of therapeutic efficacy should then establish the percentage of cases where images were judged helpful in planning management of patients, the percentage of cases where medical procedure could be avoided because of imaging findings, the number of times therapy planned before imaging changed after imaging information was obtained, or the percentage of cases in which clinicians' prospectively stated therapeutic choices changed after test information was obtained.

Evaluations in terms of patient outcome can be found in all of the retrieved models, except the one by Taube and others.[26] This can be documented in randomized clinical trials, in which specific test-treatment combinations are compared. A decision analysis comparing different diagnostic strategies may provide an investigative alternative.

A subset of authors has described a last phase, beyond the assessment of clinical effectiveness, in which cost-effectiveness and other societal effects are studied. Freedman suggested studies to monitor changes in clinical practice after the introduction of a new test.[15] In such studies, changes in diagnostic use and the frequency of test results can be documented once the new procedure is introduced into routine clinical practice. Such an evaluation can be compared with the postintroduction surveillance in the fourth phase of the evaluation of new drugs. Others proposed the assessment of societal efficacy as a final phase.[4,13,16,18,23] This phase moves beyond the individual risks and benefits of a test to an appraisal of the use of resources and medical benefits on a societal level.

DISCUSSION

In a phased evaluation strategy, more elaborate and therefore more expensive types of studies are only performed if sufficient evidence has been obtained in previous steps of the evaluation process. In this review, we identified 31 proposals for a hierarchical model of evidence or a phased evaluation scheme for medical tests. We are aware that our review has its limitations, as we only searched papers in journals and did not look systematically for proposals described in books only. Because of poor indexing, we may not have been able to identify all existing schemes.

The variety in proposals may come as a surprise to those who are familiar with the 4 or 5 phases in drug development. Why have the phases in the clinical evaluation of drugs become so well engrained in our

thinking, and why is there more variability in evaluations of tests? One of the reasons for this difference may be the absence of a strong regulatory framework. There are no clear international standards, and there is little agreement on what evidence is required or by whom in decisions about tests.[28,29] Several authors have called for harmonization of regulatory standards internationally and for more transparency regarding the clinical evidence base for new tests. If this happens, a more standardized model may be developed in the process.

Most proposals are built on the chain of steps linking tests and outcome and can be traced back to the set of levels of efficacy identified for imaging in the 1970s. Below, we would like to present a few critical thoughts to their use as phases in the assessment of tests.

Diagnostic Accuracy

Diagnostic accuracy plays a central role in most proposals. Unfortunately, the diagnostic accuracy literature suffers from poor study design, small study samples without power calculations, and suboptimal reporting.[30–33] Design, conduct, and reporting can and should be improved.[34] Most accuracy studies focus on the test in isolation, although tests are never used in a vacuum. A number of prototypical roles of tests relative to existing ones can be distinguished: replacement, triage, or add-on.[35]

Several authors have questioned the central role of test accuracy in test evaluations.[36,37] Hunink and Krestin argued that results from accuracy studies are often too late to influence management and policy decisions, given the current rapid advances in technology.[38] Accuracy may be sufficient in providing evidence of improvement or equivalence in patient outcomes, if there is a well-defined target condition, linked to effective downstream management consequences, such as effective treatment.[39–41] Yet the pivotal position of the accuracy paradigm in the schemes identified in this review is somewhat problematic, especially whenever a new test leads to a classification in disease for which there is no clinical reference standard or when the new test is thought to be better than the current reference standard. Strategies exist to deal with cases in which the reference standard result is missing in some patients or when information can be used to build a substitute or proxy for the reference standard, but when there is no accepted reference standard, other approaches have to be used.[42]

There are other problems with a central position for diagnostic accuracy. A wide range of tests is not used for diagnosis but for other purposes, such as prognosis, prediction of treatment response, selecting therapy, or for monitoring the course of disease or the effects of treatment effect. In these situations, there is not always a reference standard available, nor is it clear how the target condition should be defined.

Diagnostic Thinking Efficacy

Because diagnostic tests are often remote from health outcome, in the short term, researchers rely on more proximate efficacy measures, such as the test's effect on clinical thinking. But studies of diagnostic thinking efficacy or therapeutic efficacy are difficult to mount. At the University of Michigan in 1972 and 1973, a group of researchers tried to measure diagnostic thinking to support the work of the ACR Efficacy Committee mentioned previously. The team collected referring physicians' diagnosis prior to and after urography and their certainty in relation to receipt of the radiological information. The change in these estimates was then transformed to log likelihood ratios.[43] The original intention was to measure the degree to which clinical management was influenced by the intravenous urogram. Unfortunately, clinicians balked at the prospect of formulating a treatment plan for a patient with, say, hematunia, who had not had a urographic contrast study.[5] Consequently, the ACR Efficacy Committee deferred all attempts to measure thinking efficacy.

Even if they could be done, are such studies also necessary? The ultimate question in decisions about testing is how much net gain from testing will there be for the patient in terms of improved treatment decisions and better health outcome.[44] Despite improvements in the methodology for measuring physician confidence, one can seriously question the validity of such studies as substitutes for improvement in patient outcome. In general, their object of study is clinician behavior, not patient outcome. A negative result in a judgment and decision-making study tells us something about the included physicians and not necessarily a great deal about the qualities of the test itself or its potential for improving health outcome. Whenever clinicians do not adjust pretest probabilities or change a management plan, we should not necessarily conclude that their failure to do so was correct. Alternatively, a confident adjustment of the probability of disease or the management plan after testing does not necessarily imply that patients are

better off. Guyatt pointed out that clinicians differ systematically in their assessment of whether a given test result contributed to management, that it may be difficult to consistently be aware of clinicians' plans before the test results are available, and that clinicians' reports of what they would do before the test result is available may differ from what they actually would have done were the technology not available.[17]

This does not imply that there is no relevance at all in studying clinicians' judgment and decision making, as patient outcome after testing will usually depend on the behavior and actions of one or more physicians. If one finds that a test does not improve patient outcome, it may be important to know that the ineffective link in the testing process is a modifiable behavior of the physician with regard to the test.

Randomized Trials

If the net gain from testing has to be expressed in terms of changes in patient outcome, one could consider jumping immediately to randomized clinical trials with patient-centered outcome measures, as Guyatt proposed.[3] Running randomized trials of tests and collecting evidence of improved patient outcome after testing have almost become synonyms in many of the proposals. Is that justified?

Randomized trials of tests are more difficult to design than randomized studies of treatment. The benefits from testing may be limited to a subset of those tested, so sample size requirements can be substantial.[45] Trials of testing need a well-defined protocol that links testing, results, and downstream decisions. It is inevitable that such trials evaluate the effectiveness of testing as well as that of downstream management. These protocols may not always mimic the way the test will ultimately be used in practice, and physician compliance with such protocols may be difficult, limiting the external validity of the trial results. All of these practical problems are challenging but not insurmountable, and trials of testing can be found in the literature.

Evidence of an improvement in health is stronger than documented accuracy, but one may not always need to conduct a randomized trial to document the benefits of testing on patient outcome. Under specific circumstances, smaller scale studies of accuracy can suffice, or noncomparative studies of testing and test combinations, or modeling.[41] Elsewhere in this issue, Lord and others offer a more complete discussion of alternatives to randomized trials of testing.[40]

A Stepwise Approach?

The 4 phases in the development of drugs and devices have shown their merit. One only proceeds to more costly or more risky evaluations if there is enough evidence from previous phases. Trials in humans only take place after they have been tested thoroughly in the laboratory on animal studies, and trials in humans precede trials in patients. Can a similarly staged model be used for the evaluation of medical tests? In the early evaluation of new markers, a phased approach definitely makes sense. In the models proposed by Pepe, Sackett, and others, the first evaluations of a marker's accuracy are designed in selected subgroups, limited in size, and only when enough evidence is gathered does one move to the more costly clinical evaluations. Should one also move cautiously through the other elements of the efficacy hierarchy, one level at a time? We do not think so. Accuracy studies are neither sufficient nor always necessary for showing improvement in patient outcomes from testing. Evaluations of physicians' judgments or their behavior are not necessary, nor can they be used as a satisfactory substitute for patient outcome.

In all fairness, the ACR committee distinguished between higher and lower levels of efficacy; they did not propose a phased evaluation of tests. Neither did Fryback and Thornbury, although their hierarchy has often been interpreted that way.[38] We do not think this is justified. The levels of efficacy should not be equated with a necessary succession of phases in the evaluation of tests, nor should they be connected with a hierarchy in study design.

More recent proposals for grading recommendations about testing, such as the GRADE approach, no longer refer to levels of evidence but distinguish grading the quality of evidence—where study design obviously matters—from ranking levels of strength for recommendations.[3] The US Preventive Services Task Force, for example, used to correlate its recommendations strongly with the research design of the most important studies, whereas nowadays it considers the evidence as a whole, using 8 steps in an analytical framework, a causal pathway linking screening or other preventive services to health outcomes.[46]

Houn and others recognized that the 4-phased model for drug development is often thought of as a linear process from idea inception to product marketing, research, and development.[21] They describe how it is actually a cyclic, repetitive process that begins with the recognition of a problem and

continues through an expansive thinking phase to experimentation, assessment, and adoption. This process may be repeated as the technology is improved or modified for new uses. The process also cycles and moves ''up the rungs'' from laboratory to applied research and, ultimately, to clinical application, and it sometimes slips back to address unanticipated problems and then advances again as those problems are resolved. Similarly, Hunink and Krestin described the linear approach as a reflection of the philosophy prevalent in the industrial period. They felt that an interwoven circular approach for the evaluation of imaging, with concurrent development, assessment, and implementation of technology, would be more appropriate.[38] The same can be said for the evaluation of tests in general.

A classification of study types and outcomes has descriptive merit in understanding the published research and the gaps in knowledge. There is also value in thoughtful considerations of the quality of the available evidence when making decisions about large-scale evaluations of testing, requiring big budgets and large numbers of participants. Yet translating levels of efficacy into a linear series of phases in evaluating tests will ultimately prove to be too restrictive and may fail to do justice to the myriad of tests and the wide range of testing purposes.

REFERENCES

1. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test: lessons from the rise and fall of dexamethasone suppression test. JAMA. 1988;259(11):1699–702.

2. Lensing AW, Hirsh J. 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in postoperative patients. Thromb Haemost. 1993;69:2–7.

3. Schünemann H, Oxman A, Brozek J, et al. Rating quality of evidence and strength of recommendations: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106–10.

4. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991;11(2):88–94.

5. Fineberg HV. Evaluation of computed tomography: achievement and challenge. AJR Am J Roentgenol. 1978;131(1):1–4.

6. Loop JW, Lusted LE. American College of Radiology diagnostic efficacy studies. AJR Am J Roentgenol. 1978;131(1):173–9.

7. Thornbury JR, Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol. 1994;162(1):1–8.

8. Thornbury JR, Fryback DG. Technology assessment: an American view. Eur J Radiol. 1992;14(2):147–56.

9. Kent DL, Larson EB. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol. 1992;27(3):245–54.

10. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clin Radiol. 1995;50(8):513–8.

11. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Med Decis Making. 1988;8(4):279–89.

12. Silverstein MD, Boland BJ. Conceptual framework for evaluating laboratory tests: case-finding in ambulatory patients. Clin Chem. 1994;40(8):1621–7.

13. Pearl WS. A hierarchical outcomes approach to test assessment. Ann Emerg Med. 1999;33(1):77–84.

14. Haddow JE, Palomaki GE. ACCE: a model process for evaluating data on emerging genetic tests. In: Khoury M, Little J, Burke W, eds. Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Oxford: Oxford University Press; 2003. p 217–33.

15. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol. 1987;60(719):1071–81.

16. Taylor CR, Elmore JG, Sun K, Inouye SK. Technology assessment in diagnostic imaging: a proposal for a phased approach to evaluating radiology research. Invest Radiol. 1993;28(2):155–61.

17. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J. 1986;134(6):587–94.

18. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. Invest Radiol. 1995;30(6):334–40.

19. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol. 2007;60(11):1116–22.

20. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem. 1990;28(12):873–9.

21. Houn F, Bright RA, Bushar HF, et al. Study design in the evaluation of breast cancer imaging technologies. Acad Radiol. 2000;7(9):684–92.

22. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. Acad Radiol. 2000;7(9):681–3.

23. Zweig MH, Robertson EA. Why we need better test evaluations. Clin Chem. 1982;28(6):1272–6.

24. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002;324(7336):539–41.

25. Pepe MS. Evaluating technologies for classification and prediction in medicine. Stat Med. 2005;24(24):3687–96.

26. Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. Am J Pharmacogenomics. 2005;5(6):357–64.

27. Obuchowski NA. How many observers are needed in clinical studies of medical imaging? AJR Am J Roentgenol. 2004;182(4):867–9.

28. Walley T. Evaluating laboratory diagnostic tests. BMJ. 2008;336(7644):569–70.

29. Price CP, Christenson RH. Evaluating new diagnostic technologies: perspectives in the UK and US. Clin Chem. 2008;54(9):1421–3.

30. Smidt N, Rutjes AW, van der Windt DA, et al. Quality of reporting of diagnostic accuracy studies. Radiology. 2005;235(2): 347–53.

31. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3): 189–202.

32. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999; 282(11):1061–6.

33. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ. 2006;332(7550):1127–9.

34. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1–12.

35. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089–92.

36. Feinstein AR. Misguided efforts and future challenges for research on ''diagnostic tests.'' J Epidemiol Community Health. 2002;56(5):330–2.

37. Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology. 1999;10(3):276–81.

38. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology. 2002;222(3):604–14.

39. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol. 2008;45(3):189–95.

40. Lord SJ, Irwig L, Bossuyt PMM. Evaluating new tests: when can comparative evidence of test accuracy and other intermediate outcomes be used as an alternative to randomized controlled trials. Med Decis Making. In press.

41. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144(11): 850–5.

42. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard: a review of methods. Health Technol Assess. 2007;11(50): iii, ix–51.

43. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology. 1975;114(3):561–5.

44. Fineberg HV. Computerized tomography: dilemma of health care technology. Pediatrics. 1977;59(2):147–9.

45. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet. 2000;356(9244):1844–7.

46. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21–35.

A guide for diagnostic evaluations

Rosanna W. Peeling, Peter G. Smith and Patrick M. M. Bossuyt

Abstract | Accurate diagnostic tests have a key role in patient management and the control of most infectious diseases. Unfortunately, in many developing countries, clinical care is often critically compromised by the lack of regulatory controls on the quality of these tests. The information available on the performance of a diagnostic test can be biased or flawed because of failings in the design of the studies which assessed the performance characteristics of the test. As a result, diagnostic tests are sold and used in much of the developing world without evidence of effectiveness. Misdiagnosis leading to failure to treat a serious infection or wasting expensive treatment on people who are not infected remains a serious obstacle to health.

In this supplement, we aim to provide the first in a series of simple, user-friendly operational guides on how to design and conduct evaluations of diagnostic tests for infectious diseases that are of public health importance in the developing world. Each guide will contain a set of general principles on the design and conduct of diagnostic evaluations followed by disease-specific considerations. The first in this series is the malaria guide. This article provides background information and discusses why such guides are needed and their importance in improving the diagnosis of infectious diseases in the developing world.

THE NEED FOR GOOD QUALITY DIAGNOSTIC TESTS

The lack of access to good quality diagnostic tests for infectious diseases contributes to the enormous burden of ill health in the developing world, where infectious diseases are the major causes of death and account for more than half of all deaths in children[1] (TABLE 1, FIG. 1). Each year, more than 2 million people die of malaria, approximately 4 million of acute respiratory infections and almost 3 million of enteric infections. HIV and tuberculosis together are estimated to kill some 5.8 million people each year[2,3]. More than 95% of these deaths are in developing countries. Early diagnosis and treatment not only reduces the risk of the patient developing long-term complications but for diseases such as tuberculosis, sexually transmitted infections (STIs) and HIV, prompt treatment also reduces further transmission of the disease to other members of the community.

A confident diagnosis can sometimes be made on the basis of clinical signs or symptoms but accurate diagnosis usually requires a specific diagnostic test, often involving access to a diagnostic laboratory. In settings where access to diagnostic laboratory services is limited, the WHO recommends the use of a syndromic approach to clinical management, where patients presenting with a particular syndrome are treated for all of the major causes of the syndrome. Algorithms for syndromic management have been developed for STIs and for common childhood diseases, the latter through the integrated management of childhood illness (IMCI)[4,5]. Although such algorithms are simple to use and the recommended treatment packages are generally inexpensive, a major disadvantage of this approach is the risk of giving inappropriate treatment to people without the syndromically diagnosed disease and the accompanying potential for inducing antibiotic resistance. Diagnostic tests can complement syndromic management by facilitating evidence-based management of patients, improving the specificity of treatment and, in some diseases, allowing contact tracing and other disease-control measures.

Laboratory testing is perhaps most useful for detection of asymptomatic infections to prevent development of sequelae and transmission, and for public health surveillance and interventions. TABLE 2 shows the role of diagnostic tests in the control of some of the diseases that are prevalent in developing countries[6].

Good quality diagnostic tests that are fit for purpose and provide accurate results are therefore of paramount importance in reducing the burden of infectious diseases (BOX 1). The choice of which diagnostic test to use depends on which tests have been approved for use by regulatory authorities in a particular country (if they are regulated at all) and which tests have been purchased for use in the health service; and the physician's decision on which of the available tests he or she judges might be useful in clinical decision making. Unfortunately, in many developing countries, clinical care is often critically compromised by the lack of regulatory controls on the quality of diagnostics, and physicians can be faced with having to select tests based only on information provided in the product insert or on published data that often originate from inadequate or flawed study designs.

Table 1 | **Top five causes of deaths in selected regions in 2001**

	Sub-Saharan Africa	South Asia	Europe and central Asia
1	HIV/AIDS	Ischaemic heart disease	Ischaemic heart disease
2	Malaria	Lower respiratory infections	Cerebrovascular disease
3	Lower respiratory infections	Perinatal conditions	Lung cancer
4	Diarrhoeal diseases	Cerebrovascular disease	Chronic obstructive pulmonary disease
5	Perinatal conditions	Diarrhoeal diseases	Self-inflicted injuries

Data taken from REF. 22.

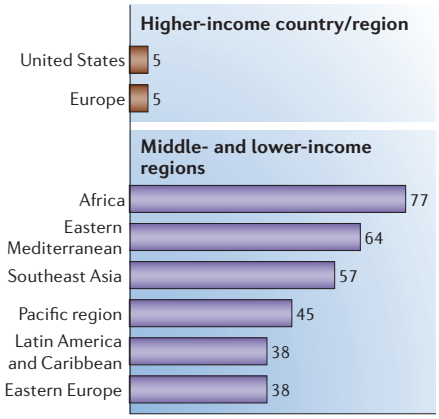


Figure 1 | **Childhood deaths from infectious diseases.** The graph shows the percentage of childhood deaths that were attributable to infectious diseases in selected world regions for the period 2000–2003. In the higher-income section, Europe excludes Eastern Europe. Data taken from REF. 1.

THE DEVELOPMENT OF DIAGNOSTIC TESTS

The development of a diagnostic test usually follows a path from identification of the diagnostic target and optimization of test reagents to the development of a test prototype (FIG. 2). Proof-of-principle studies are then conducted to establish that the test detects the intended target. The test then undergoes further evaluations, first using 'convenience' samples or archived specimens, followed by evaluations in populations of intended use. These trial results are used to obtain data for regulatory submission and approval so that the tests can be marketed and sold in a country. For post-approval marketing purposes, companies often fund physicians to conduct studies to demonstrate the utility and potential impact of the diagnostic test.

TEST CHARACTERISTICS REQUIRING EVALUATION

Diagnostic tests can be purchased by patients, health providers, clinics, hospitals, national disease control programmes, procurement agencies for organizations or donors. Although the criteria on which procurement decisions are made can vary, selections are generally based on the factors discussed below.

Test performance. Test sensitivity and specificity, and the positive and negative predictive values of a test, are important considerations. High sensitivity is important for a screening test for diseases such as syphilis where a missed diagnosis has serious consequences. Poor specificity might matter less if over-treatment rarely results in adverse side effects, as in the treatment for syphilis, but might be a serious disadvantage if the treatment is highly toxic as, for example, is the case with drugs used to treat advanced trypanosomiasis (sleeping sickness).

Ease of use. The number of processing steps, whether the test can use whole blood and the need for accurate timing will influence the extent of training and supervision required.

Conditions of use. In hot or humid conditions, the selection of tests that are heat-stable and individually packaged in moisture-proof pouches is a priority.

Conditions of storage. There are defined storage temperatures for most tests. If the temperature in the clinic is above 30°C and the accuracy of the test results is not guaranteed above this temperature, periodic quality-control checks to ensure the ongoing validity of the tests are needed.

Shelf life. A long shelf life reduces the pressure on the supply chain and the probability of wastage of expired tests. Tests which have a shelf life in excess of 18 months are recommended for use in remote, poorly resourced areas.

Of these factors, the test performance is of paramount importance. Many diagnostic evaluations therefore focus primarily on evaluations of test performance, that is, the sensitivity and specificity or the positive and negative predictive values.

LACK OF REGULATORY STANDARDS AND GUIDELINES

National regulatory processes should provide safeguards for the safety and effectiveness of drugs used in a country. The tightening of governmental regulatory requirements for drugs in developing countries has done much to improve the standardization and quality of drug trials, in which efficacy and adverse effects are

Table 2 | **Examples of the role of diagnostic tests in the control of some of the major infectious diseases in developing countries***

Disease	Control strategy	Diagnostic test(s)	Role of diagnostic tests		
			Case management	Screening	Surveillance
Acute respiratory infections	Vaccination and syndromic case management	Diagnosis based on syndromes	No	No	Limited use
Diarrhoeal diseases	Vaccination and syndromic case management	Stool culture	No	No	Yes, for cholera
Malaria	Vector control and case management	Blood film	Yes	No	Yes, to identify outbreaks outside hyper-endemic areas
		Serology	Not usually indicated	Yes, at blood banks	No
HIV	Health promotion, STI control, voluntary counselling and testing, PMTCT	Serology	Yes	Yes, at blood banks and pregnant women for PMTCT	Yes, in sentinel surveillance in defined population groups
		$CD4^+$ T-cell count and viral load	Yes (where ARVs are available)	No	No
Tuberculosis	Case detection followed by DOTS	Sputum microscopy (and culture)	Yes	No	Yes, in national reporting schemes
		Tuberculin skin test	Of debatable use	Yes	No

*Data taken from REF. 6. ARVs, antiretrovirals; DOTS, directly observed therapy strategy; PMTCT, prevention of mother-to-child transmission of HIV; STI, sexually transmitted infection.

Box 1 | **Role of diagnostic tests in the management and control of infectious diseases**

Diagnostic tests have a crucial role in the management of patients and in the control of infectious diseases. Their uses can include some or all of the following:

Patient management
Diagnostic tests can support clinical decision making, especially when the clinical symptoms are not specific enough to allow diagnosis of a specific infection, as is often the case. They can be used to confirm or rule out a clinical diagnosis in symptomatic patients, an example being the use of sputum microscopy to diagnose tuberculosis in patients presenting with a cough.

Screening for asymptomatic infections
Many infectious diseases cause non-specific symptoms or no symptoms at all. Undetected and untreated infections can lead to serious long-term complications, can continue the chain of transmission and, for some infections, can cause adverse fetal outcomes in pregnant women. Tests to screen individuals for such asymptomatic infections might prevent clinical disease and stop or reduce disease transmission in the community. In the case of syphilis, tests can avert stillbirths and congenital syphilis.

Surveillance, including verification of elimination
Surveillance is the cornerstone of successful disease control or elimination programmes as it enables programme managers to monitor the effectiveness of intervention strategies and can help to identify populations that require continuing interventions. For some infections, such as smallpox and measles, the clinical features are distinctive and surveillance can be based on clinical findings alone. For others, for example polio, the clinical diagnosis of acute flaccid paralysis is strongly suggestive, but there are other causes of this syndrome. As the incidence of polio falls, the proportion of cases of flaccid paralysis owing to polio decreases and it might be necessary to confirm cases by a laboratory test, such as electron microscopy of a faecal sample. The clinical features of many other important infectious diseases are insufficiently distinctive for surveillance purposes, especially in the early stages. For these diseases diagnostic tests are required.

Epidemiological studies
Diagnostic tools can be important in rapid assessments of the disease burden in particular populations to allow the rational design of control strategies. This can be especially important in outbreak investigations.

Detection of infections with markers of drug resistance
The rise of drug resistance renders many disease control programmes ineffective. For malaria and tuberculosis, the development of drug resistance has led to more costly and complex drug regimens. The use of diagnostic tests for drug resistance surveillance is fundamental to the refinement of treatment strategies and the allocation of scarce resources.

assessed and compared. Unfortunately, regulatory standards are often lacking for diagnostic tests, especially those targeting diseases that are uncommon in industrialized countries. As a result, diagnostic tests are often sold in the developing world without any formal evaluation of their performance and effectiveness. An exception to this is tests used for blood banking, for which rigorous international standards exist.

WHO/TDR conducted a global survey of regulatory practices for diagnostic tests in 2001. A questionnaire was sent to all 191 WHO member states to enquire whether *in vitro* diagnostics, other than those used for blood banking, were regulated in their country and, if so, whether clinical trials were required for regulatory approval. Of the 85 countries that responded, less than half (48%) reported that they regulated *in vitro* diagnostics for infectious diseases[7]. A greater number of countries in the developed world regulate *in vitro* diagnostics compared with the number in the developing world (FIG. 3a). Of the countries that regulated diagnostics, 68% required the submission of clinical trial data (FIG. 3b).

There is also variability from country to country in terms of which tests for specific infectious diseases are regulated. Of the 24 countries that provide these data, 83% regulated diagnostics for HIV, 92% for hepatitis, 42% for STIs and 13% each for tuberculosis and malaria[7].

An industry survey conducted by WHO/TDR in 2003 found that companies can spend from as little as US$2,000 to more than US$1,000,000 on diagnostic trials of different products, with some diagnostic trials conducted in as few as 15 patients (unpublished TDR data).

Even when clinical trials are mandated by regulatory authorities, there is a lack of national and international guidelines for the evaluation of diagnostic tests for diseases that are prevalent in developing countries. Standards for the evaluation of diagnostic tests are set by regulatory bodies such as the US Food and Drugs Administration (FDA) and the European Union, and, for example, the Clinical and Laboratory Standards Institute in the USA publishes standards that are widely used by manufacturers targeting markets in established economies. However, these standards were developed for the evaluation of tests in developed countries and are often not applicable for diseases that are prevalent in the developing world.

Data from clinical trials designed to evaluate the performance characteristics of diagnostic tests are often found on product inserts or they remain in the company files. Although every product insert contains claims of high sensitivity and specificity, there is no requirement to report the sample size or the confidence intervals. One product dossier recently submitted to WHO/TDR showed that the test was evaluated in more than 100 patients, of whom only three were positive for the disease by the reference standard (unpublished TDR data); the claim was that the test is 100% sensitive and 100% specific. In many countries the lack of regulatory oversight on the design and conduct of diagnostic evaluations has led to inflated claims of test performance in product inserts.

This underscores the need for a set of international standards to regulate diagnostics for infectious diseases (outside of blood banking). A global harmonization task force has published guidance on the regulation of medical devices and a scheme for classifying medical devices, but plans for international standards for regulatory approval of diagnostic tests of public health importance in the developing world are still in the distant future.

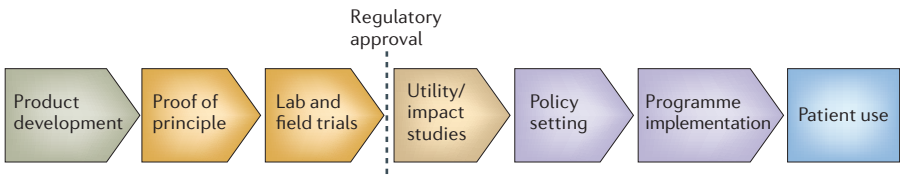


Figure 2 | **The bench-to-bedside pathway of diagnostics development and evaluation.** The development of a diagnostic test usually follows a path from identification of the diagnostic target and optimization of test reagents to the development of a test prototype that then undergoes a series of evaluations. Reproduced with permission from REF. 13.

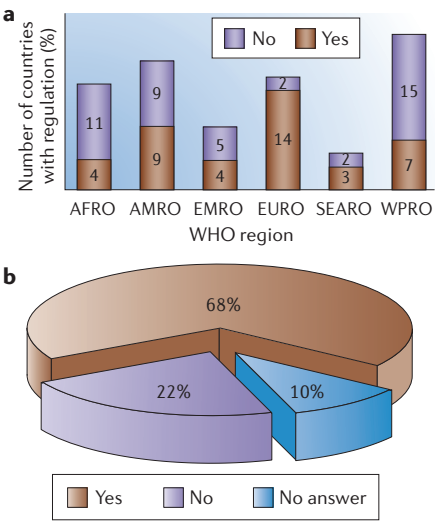


Figure 3 | **Regulation of diagnostics. a** | The number of countries that regulate diagnostics by region. The WHO regions are as follows: AFRO, Africa; AMRO, Americas; EMRO, East Mediterranean; EURO, Europe; SEARO, South and Southeast Asia and WPRO, Western Pacific. **b** | The number of countries that require clinical evaluation of diagnostics for regulatory approval.

EVALUATIONS PUBLISHED IN THE PEER-REVIEWED LITERATURE

The design and quality of trials of diagnostic tests have profound effects on the estimation of the performance characteristics of these tests. Trials to evaluate the performance and operational characteristics of diagnostic tests can be conducted by test manufacturers, public health agencies and end-users such as physicians and laboratory managers in hospitals and clinics. These studies are sometimes published in the peer-reviewed literature, either sponsored by the manufacturing company or conducted by independent investigators. Reviews of diagnostic publications since the late 1970s have shown that although the quality of diagnostics trials is improving, many are still lacking in rigour[8–12]. For industry-sponsored studies, this might be because diagnostics for infectious diseases that are prevalent in the developing world tend to be produced by small biotechnology companies that have relatively few resources and limited expertise in field trials[13]. Some common design problems in diagnostic evaluations are listed below.

Evaluation in an inappropriate study group. To assess properly how a test will perform in routine use, diagnostic evaluations must be performed in a study group that is sampled from the population for which the test is intended. The diagnostic performance can vary in symptomatic and asymptomatic patients, and can also differ when diagnostics are used to detect active versus latent disease. The data from diagnostic evaluations are only useful if there is an adequate description of the study group used in the evaluation, with well-defined inclusion and exclusion criteria and an adequate sample size for each sub-population.

Evaluation in an inappropriate setting. Evaluations in low-prevalence settings can result in a much higher proportion of false-positive to true-positive results than would be found in a high-prevalence setting.

Inappropriate purpose. Diagnostic tests used for screening of asymptomatic patients, diagnosis in symptomatic patients, surveillance, and verification of elimination all require different performance characteristics. Diagnostic trials should be designed and conducted for a specific purpose to yield meaningful results.

Inappropriate reference standard test. The reference standard test is the comparator for the test under evaluation. The selection and the quality of the reference standard test directly affect the measurement of test performance. An ongoing challenge for diagnostic evaluations is to deal with trials where the test under evaluation is more sensitive and/or specific than the reference standard test.

Inadequate sample size. For reasons of economy and time, diagnostic evaluations are often conducted in a small number of patients, leading to wide confidence intervals around the estimates of sensitivity or specificity[14].

Lack of blinding. Providing readers of the reference standard test with the results from the test under evaluation, or vice versa, might artificially inflate the agreement between both. A recent review of studies reporting diagnostic evaluation of tests for tuberculosis showed that only 34% reported any form of blinding[12].

The quality of evaluation trials. The proficiency of the site staff in performing the reference standard test and the test under evaluation is often difficult to discern from publications or from manufacturers' dossiers.

Reid *et al.* examined diagnostic evaluations reported in four prominent general medical journals from 1978 to 1993 and found that less than half the studies fulfilled more than three of the seven methodological standards outlined in FIGURE 4 (REF. 10; see also REF. 15 for a more recent evaluation).

INITIATIVES TO IMPROVE THE STANDARD OF DIAGNOSTIC EVALUATIONS

Apart from the standards set by national regulatory agencies such as the US FDA or the equivalent organization in Thailand, there are several other initiatives that provide guidelines on diagnostic evaluations.

Meta-analysis and systematic reviews. In 1994, guidelines were published for the conduct, reporting and critical appraisal of meta-analyses evaluating diagnostic tests[16]. A systematic review of near-patient test evaluations in primary care was conducted in 1999 in an attempt to identify and synthesize results from studies that examined the performance and effect of such tests[17]. One hundred and one relevant papers published between 1986 and 1996 were identified. The authors concluded that the quality of the papers was generally low. The performance of most tests had not been adequately evaluated and most papers reported biased assessments of the effect of near-patient tests on patient outcomes, organizational outcomes or cost.

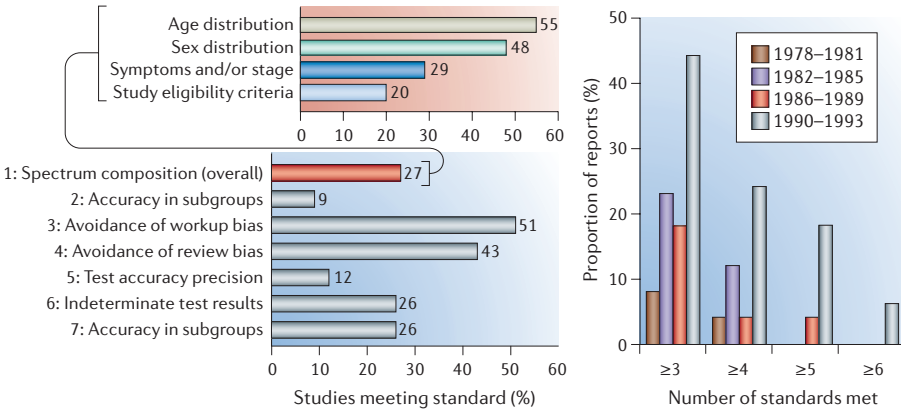


Figure 4 | **Proportion of diagnostic evaluations meeting accepted standards.** The seven standards are shown on the left. The data are taken from REF. 10.

Such meta-analyses serve both to provide an overall summary of diagnostic accuracy from several studies and to identify deficits in published studies that need to be addressed in future studies.

The STARD initiative. The variable quality of publications on diagnostic evaluations led to the launch of the STARD (Standards for Reporting of Diagnostics Accuracy) initiative in 2003 (REFS 18,19). STARD aims to improve the quality of diagnostic test evaluation reported in the peer-reviewed literature. The STARD checklist is included in the general guide (*Evaluation of diagnostic tests for infectious diseases: general principles*) in this supplement. It is hoped that this initiative will gradually have an impact on the design and execution of trials in the developing world.

Other reference material. Other available reference material ranges from articles on how to read a paper on diagnostics to books on the design and conduct of field trials of health interventions in developing countries[20,21]. Although general recommendations can be found in various sources, there is limited disease-specific guidance on the design and conduct of diagnostic trials for diseases prevalent in the developing world. These disease-specific considerations include which populations should be targeted, what reference standard should be used, how to define case and control populations, how sampling should be performed and how to ensure blinding of results.

DEEP AND THE DEVELOPMENT OF BEST PRACTICE

In the absence of robust standards for diagnostic trials, scarce public sector resources might be wasted on diagnostics that not only lead to mismanagement of patients but also have little impact on reducing the disease burden. There is a need for stricter controls on the introduction and use of diagnostic tests in national public health programmes in many developing countries, based on the rigorous evaluation of tests before, or during, deployment. Data on the performance and operational characteristics of diagnostic tests from well-designed trials are required to allow those responsible for procuring tests to make informed decisions about the choice of specific tests.

WHO/TDR has assembled a Diagnostic Evaluation Expert Panel (DEEP) to advise WHO/TDR and its close collaborator, the Foundation for Innovative New Diagnostics (FIND) (BOX 2), on recommendations for best practice in the design and conduct of diagnostic trials for selected infectious diseases of public health importance in the developing world. One of the first tasks of the panel was to produce a set of general principles for the design and conduct of diagnostic evaluations that are harmonized with the current standards established by national and international agencies and, by various initiatives, to improve the standard of diagnostic evaluations. This will be followed by a series of disease-specific recommendations on how the necessary methodological standards can be fulfilled in the evaluation of diagnostics for diseases of public health importance to the developing world. The first in this series is the malaria guide.

Our aim is to provide a set of simple, user-friendly operational guidelines on the design and conduct of diagnostic trials, to support regulatory agencies in the consideration of registration applications, to provide procurement agencies and international health agencies with performance benchmarks, and to enable scientists in developing countries, especially those working on disease control in the public sector, to evaluate diagnostic tests in accordance with international standards.

Box 2 | **TDR and FIND**

The UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR) was established in 1975 to conduct research aimed at the development of new tools for the control of tropical diseases, and to train researchers from disease-endemic countries. The tools, which include drugs, vaccines and diagnostics, are developed through public–private partnerships. Current TDR priorities for diagnostics include visceral leishmaniasis, African trypanosomiasis, schistosomiasis, malaria, dengue and tuberculosis.

Recognizing that the biotechnology revolution of past decades had not resulted in significant changes in diagnostic practices in disease-endemic countries, a new, independent, not-for-profit entity, the Foundation for Innovative New Diagnostics (FIND), was created in 2003 to respond specifically to the need for better diagnostics for the developing world. FIND works in close collaboration with the diagnostics programme in TDR to overcome the obstacles that have blocked academic, government and corporate entities from moving promising ideas through the developmental pipeline and ensuring their uptake by public health systems to decrease global health inequities.

Rosanna W. Peeling is at the UNICEF/UNDP/World Bank/WHO Special Programme for Research & Training in Tropical Diseases (TDR), World Health Organization, 20 Avenue Appia, CH-1211 Geneva 27, Switzerland.*

Peter G. Smith is at the Infectious Diseases Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Patrick M. M. Bossuyt is at the Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands.

**e-mail: peelingr@who.int*

doi:10.1038/nrmicro1522

1. Kent, M. M. & Yin, S. *Controlling Infectious Diseases* (Population Reference Bureau, Washington DC, 2006).
2. WHO. *World Health Report* [online], <http://www.who.int/whr/2005/whr2005_en.pdf> (2005).
3. WHO/UNAIDS. *AIDS Epidemic Update* [online], <http://data.unaids.org/Publications/IRC-pub06/epi_update2005_en.pdf> (2005).
4. WHO. *Guidelines for the Management of Sexually Transmitted Infections* [online], <http://www.who.int/reproductive-health/publications/rhr_01_10_mngt_stis/guidelines_mngt_stis.pdf> (2001).
5. WHO. *Integrated Management of Childhood Illness Information Pack* [online], <http://www.who.int/child-adolescent-health/publications/IMCI/WHO_CHS_CAH_98.1.htm> (1998).
6. Mabey, D., Peeling, R.W., Ustianowski, A. & Perkins, M. Diagnostics for the developing world. *Nature Rev. Microbiol.* **2**, 231–240 (2004).
7. Cunningham, J. *et al.* in *Diagnostics for Tuberculosis: Global Demand and Market Potential* 116–120 (WHO, in the press).
8. Ransohoff, D.F. & Feinstein, A. R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New Engl. J. Med.* **299**, 926–930 (1978).
9. Lach, M. S. *et al.* Spectrum bias in the evaluation of diagnostic tests: lessons learnt from the rapid dipstick test for urinary tract infections. *Ann. Intern. Med.* **117**, 135–140 (1992).
10. Reid, M. C., Lachs, M.S. & Feinstein, A. Use of methodological standards in diagnostic test research. Getting better but still not good. *J. Amer. Med. Assoc.* **274**, 645–651 (1995).
11. Small, P. M. &Perkins, M. D. More rigour needed in trials of new diagnostic agents for tuberculosis. *Lancet* **356**, 1048–1049 (2000).
12. Pai, M. & O'Brien, R. Tuberculosis diagnostic trials: do they lack methodological rigour? *Expert Rev. Mol. Diag.* **6**, 1–6 (2006).
13. Kettler, H., White, K. & Hawkes, S. *Mapping the Landscape of Diagnostics for Sexually Transmitted Infections* [online], <http://www.who.int/tdr/publications/publications/pdf/mapping_landscape.pdf> (2004).
14. Bachmann, L. M., Puhan, M. A., ter Riet, G. & Bossuyt, P. M. Sample size of studies on diagnostic accuracy: literature survey. *Br. Med. J.* **332**, 1127–1129 (2006).
15. Smidt, N. *et al.* Quality of reporting diagnostic accuracy studies. *Radiology* **235**, 347–353 (2005).
16. Irwig, L. *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Intern. Med.* **120**, 667–676 (1994).
17. Delaney, B. C. *et al.* Systematic review of near-patient test evaluations in primary care. *Br. Med. J* **319**, 824–827 (1999).
18. Bossuyt, P. M. *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Clin. Chem.* **49**, 1–6 (2003).
19. Bossuyt, P. M. *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin. Chem.* **49**, 7–18 (2003).
20. Delaney, B. C., Wilson, S., Fitzmaurice, D., Hyde, C. & Hobbs, R. Near-patient tests in primary care:setting the standards for evaluation. *J. Health Serv. Res. Policy* **5**, 37–41 (2000).
21. Smith, P.G. & Morrow, R. H., eds *Field Trials of Health Interventions in Developing Countries: A Toolbox*. (Macmillan, London, 1996).
22. Mathers, C. D., Lopez, A. D. & Murray, C. L. in *Global Burden of Disease and Risk Factors* (Lopez, A. D. *et al.*, eds) 45–93 (Oxford Univ. Press, 2006).

Evaluation of diagnostic tests for infectious diseases: general principles

The TDR Diagnostics Evaluation Expert Panel

I. INTRODUCTION

A diagnostic test for an infectious agent can be used to demonstrate the presence or absence of infection, or to detect evidence of a previous infection (for example, the presence of antibodies). Demonstrating the presence of the infecting organism, or a surrogate marker of infection, is often crucial for effective clinical management and for selecting other appropriate disease control activities such as contact tracing. To be useful, diagnostic methods must be accurate, simple and affordable for the population for which they are intended. They must also provide a result in time to institute effective control measures, particularly treatment. For some infections, early diagnosis and treatment can have an important role in preventing the development of long-term complications or in interrupting transmission of the infectious agent. In a broader context, diagnostic tests can have multiple uses, including: patient management, especially when clinical symptoms are not specific for a particular infection (as is often the case); screening for asymptomatic infections; surveillance; epidemiological studies (for example, rapid assessments of disease burden or outbreak investigations); evaluating the effectiveness of interventions, including verification of elimination; and detecting infections with markers of drug resistance.

Recent technological developments have led to the proliferation of new, rapid diagnostic tests that hold promise for the improved management and control of infectious diseases. Whether these tests are useful in a given setting and, if so, which test is most appropriate are questions that can be answered only through evaluations in the appropriate laboratory, clinical or field settings.

Many variables can influence the performance of tests in different settings. These include differences in the characteristics of the population or the infectious agent, including the infection prevalence and genetic variation of the pathogen or host, as well as the test methodology — for example, the use of recombinant or native antigen or antibody, whether the test is manual or automatic, the physical format of the test and local diagnostic practice and skills. Therefore, wherever possible, test evaluations should be performed under the range of conditions in which they are likely to be used in practice. In some situations, such evaluations can be facilitated through multi-centre trials.

Lack of resources and expertise limit the ability of many developing countries to perform adequate evaluations of diagnostic tests, and many new tests are marketed directly to end-users who lack the ability to assess their performance. The onus is therefore on those who perform the evaluations to ensure that the quality of the methods and the documentation used is such that the findings add usefully to the pool of knowledge on which others can draw. The Standards for Reporting of Diagnostic Accuracy (STARD) initiative has developed a sequenced checklist to help to ensure that all relevant information is included when the results of studies on diagnostic accuracy are reported[1–4] (APPENDIX 1).

Evaluations of diagnostic tests must be planned with respect to their use for a clearly defined purpose, carefully and systematically executed, and must be reported in a way that allows the reader to understand the study methods and the limitations involved and to interpret the results correctly. This will help to avoid the financial and human costs associated with incorrect diagnoses, which can include poor patient care, unnecessary complications, suffering and, in some circumstances, even death.

This document is concerned with general principles in the design and conduct of trials to evaluate diagnostic tests. It is not a detailed operational manual and should be used alongside detailed descriptions of statistical methods, clinical trial guides and other reference materials given in the reference list.

The goals of this document are to facilitate the setting of appropriate standards for test evaluation; to provide best-practice guidelines for assessing the performance and operational characteristics of diagnostic tests for infectious diseases in populations in which the tests are intended to be used; to help those designing evaluations at all levels, from test manufacturers to end-users; and to facilitate critical review of published and unpublished evaluations, with a view to selecting or approving tests that have been appropriately evaluated and shown to meet defined performance targets. The target audience for this document includes institutions and research groups that are planning trials of diagnostic tests; organizations that fund or conduct trials of diagnostic tests; agencies responsible for the procurement of diagnostic tests; diagnostic test manufacturers; and regulatory authorities.

II. CHARACTERISTICS ASSESSED IN EVALUATIONS OF DIAGNOSTIC ACCURACY

1. Performance characteristics

The basic performance characteristics of a test designed to distinguish infected from uninfected individuals are sensitivity, that is, the probability that a truly infected individual will test positive, and specificity, that

is, the probability that a truly uninfected individual will test negative. These measures are usually expressed as a percentage.

Sensitivity and specificity are usually determined against a reference standard test, sometimes referred to as a 'gold standard' test, that is used to identify which subjects are truly infected and which are uninfected. Errors in measuring the sensitivity and specificity of a test will arise if the 'gold standard' test itself does not have 100% sensitivity and 100% specificity, which is not infrequently the case. Evaluating a diagnostic test is particularly challenging when there is no recognized reference standard test.

Two other important measures of test performance are positive predictive value (PPV), the probability that those testing positive by the test are truly infected, and negative predictive value (NPV), the probability that those testing negative by the test are truly uninfected. Both of these measures are often expressed as percentages. PPV and NPV depend not only on the sensitivity and specificity of the test, but also on the prevalence of infection in the population studied (BOX 1). The reproducibility of a test is an assessment of the extent to which the same tester achieves the same results on repeated testing of the same samples, or the extent to which different testers achieve the same results on the same samples, and is measured by the percentage of times the same results are obtained when the test is used repeatedly on the same specimens. Reproducibility can therefore be measured between operators or with the same operator, or using different lots of the same test reagent. The accuracy of a test is sometimes used as an overall measure of its performance and is defined as the percentage of individuals for whom both the test and the reference standard give the same result (that is, the percentage of individuals whom both tests classify as infected or uninfected). Note that the use of this measure of diagnostic accuracy is of limited value and is often difficult to interpret, as it depends on sensitivity, specificity and the prevalence of infection.

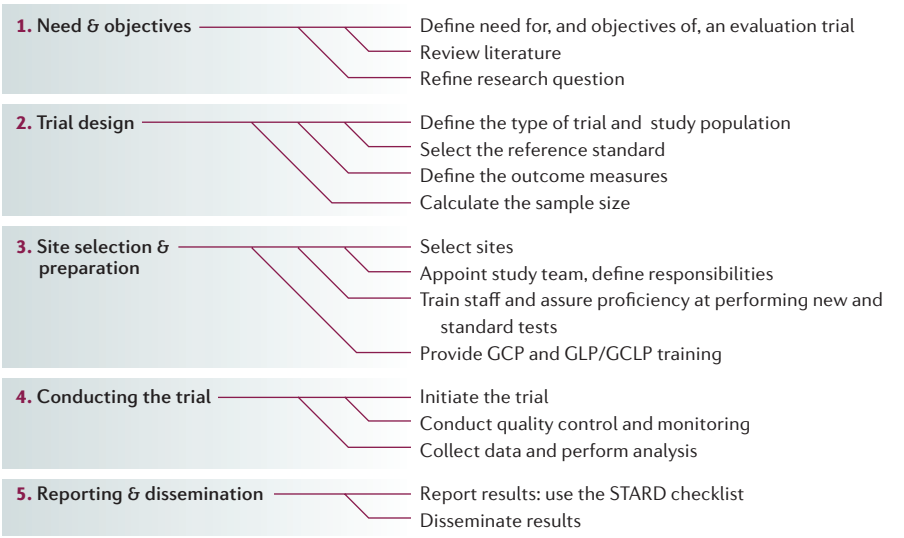


Figure 1 | **Essential elements in designing diagnostic test evaluations.** GCP, good clinical practice; GCLP, good clinical laboratory practice; GLP, good laboratory practice; STARD, standards for reporting of diagnostic accuracy. See Section III, 2.13.

2. Operational characteristics

Operational characteristics include the time taken to perform the test, its technical simplicity or ease of use, user acceptability and the stability of the test under user conditions. The ease of use will depend on the ease of acquiring and maintaining the equipment required to perform the test, how difficult it is to train staff to use the test and to interpret the results of the test correctly, and the stability of the test under the expected conditions of use. All of these characteristics are important for determining the settings in which a diagnostic test can be used and the level of staff training required. Information on test stability — how tests can be stored in peripheral healthcare settings and for how long — is crucial for decisions on procurement.

III. ESSENTIAL ELEMENTS IN THE DESIGN OF DIAGNOSTIC TEST EVALUATIONS

The design of a study is likely to be greatly improved if this process is approached systematically along the lines outlined in FIGURE 1.

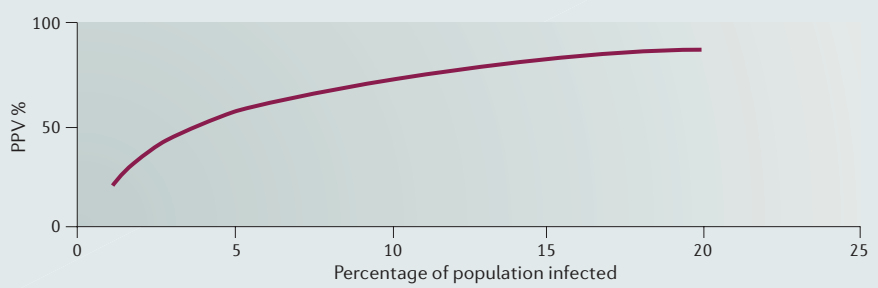
1. Defining the need for a trial and the trial objectives

Before conducting a trial, the specific need for the trial should be assessed. The purpose of the study and the degree to which the outcome is likely to contribute to improved health outcomes and/or further knowledge about the performance of the test should be specified.

First, the problem must be identified. Examples include the need for a screening test because an infection is asymptomatic but undetected infection can cause serious complications, as is the case for screening pregnant women to prevent congenital syphilis; the need for a rapid point-of-care test because of low patient return for results of laboratory-based tests that require return visits; and the need for a test that can be performed on specimens obtained through less-invasive procedures, such as a blood test instead of a lumbar puncture for determining the stage of disease in African trypanosomiasis.

Box 1 | **Dependence of PPV on prevalence**

The positive predictive value (PPV) of a test will depend not only on the sensitivity of the test but also on the prevalence of the condition within the population being tested. The figure below shows how the positive predictive value for a test with 96% sensitivity varies according to the prevalence of infection in the population.



The purpose for which the new test is designed, for example, whether the test is to be used for case management, to screen asymptomatic infections, for surveillance or to verify elimination, must also be defined. This should include defining the objectives of the evaluation — both the overall objective, for example, improving the quality of diagnosis for patient management or surveillance, and the specific objectives, for example, assessing test performance, acceptability, the impact on patient care or the prevention of complications.

Finally, the relevant literature must be reviewed and the research question refined. Planning a study should include a comprehensive review of the relevant literature for the diagnostic test under evaluation and other relevant tests for the infection under study, including an assessment of the strengths and limitations of previous studies. Where possible, manufacturers' clinical and analytical data for the new test should be assessed. The outcome of a review of previous work should inform assessment of the need for another trial and the specific areas in which further information is needed. As a consequence of the review, the research question might be refined.

2. General considerations in the design of evaluation trials

The design of the evaluation trial will depend on its purpose and the population for which the test is intended. See REF. 5 for a general discussion of the factors to take into account when planning field trials. In general, a diagnostic test should be evaluated using methods and equipment that are appropriate for that purpose. The staff performing the evaluation should be appropriately trained so that they are proficient in performing the test being evaluated and the comparator tests.

2.1. Defining the study population. There should be clear specification of the eventual target population in which the diagnostic test will be used. Defining the target population must take into account the probable purpose of the test. For example, will it replace an existing test, will it be used as a triage instrument to identify those in need of further investigation, or will it be used as an additional test in a diagnostic strategy for case finding or for screening asymptomatic individuals? The actions that are to be guided by the use of the test, such as starting or withholding treatment, must also be considered. It is of little value to evaluate tests that are unlikely to be affordable or accessible to the target population, or which yield results that are unlikely to influence patient care or public health practice.

2.2. Subjects to be included in the study. Two common circumstances in which diagnostic tests are deployed are:

a Screening people presenting to a clinic who have symptoms that might be caused by the infection to identify those who are truly infected (for example, persons presenting with a fever that might be caused by malaria).

b Distinguishing infected people from non-infected people in a population, irrespective of whether or not they have any symptoms that might be characteristic of the infection.

Generally, in situation (a), tests with high sensitivity will be required so that a high proportion of all truly infected patients are identified for treatment. In situation (b), if the infection is rare, high specificity will be required or else a high proportion of those who test positive could be false positives (that is, the test will have a poor PPV). In either circumstance it is necessary to identify a group of truly infected and truly uninfected individuals to assess sensitivity and specificity, respectively.

For situation (a), a common design for an evaluation study is to enroll consecutive subjects who are clinically suspected of having the target condition. The suspicion of infection can be based on presenting symptoms or on a referral by another healthcare professional. These participants then undergo the test under evaluation as well as the reference standard test. In studies in which only a small proportion of those tested are likely to be infected, all subjects can be subjected to the reference standard test first. All positives and only a random sample of test negatives can then be subjected to the test under evaluation. This can lead to more efficient use of resources if the target condition is rare.

Tests can sometimes be evaluated using stored specimens collected from those with known infection status. Such studies are rapid and can be of great value but there is a risk that they can lead to inflated estimates of diagnostic accuracy (when the stored samples have been collected from the 'sickest of the sick' and the 'healthiest of the well'). The estimate of specificity can also be biased if, for example, the negative samples relate only to a group with one alternative condition, rather than a group including the full range of conditions that can present with symptoms that are similar to the infection under study.

Box 2 | **Multi-centre studies**

Advantages
- Larger sample size
- Representative of more than one population so findings are more generally applicable

Disadvantages
- Greater expense
- Quality control and coordination more difficult
- Site conditions might not be comparable

2.3. The study setting. The setting where patients or specimens will be recruited and where the evaluation will be conducted should be defined. This might be in a clinic or laboratory, a remote health post or a hospital. Tests will probably perform differently in a primary care setting compared with a secondary or tertiary care setting. The spectrum of endemic infections and the range of other conditions observed can vary from setting to setting, depending on the referral mechanism. Other factors that can affect test performance and differ between sites include climate, host genetics and the local strains of pathogens. Because the test characteristics can vary in different settings, it is often valuable to consider conducting multi-centre studies. Some of the advantages and disadvantages of multi-centre studies are shown in BOX 2.

2.4. Retrospective and prospective evaluations. Diagnostic evaluations can be performed both retrospectively, using well-characterized archived specimens, and prospectively, using fresh specimens. The choice depends on the availability of appropriate specimens and whether the research question can be answered wholly or in part using archived specimens. Some advantages and disadvantages of using archived specimens are shown in BOX 3.

2.5. Eligibility criteria. The eligibility criteria are generally defined by the choice of the target population, but additional exclusion criteria can be used for reasons of safety or feasibility. The researcher must consider, for example, whether or not patients with co-morbidity or other conditions likely to influence the study results will be excluded. For infectious diseases, additional exclusion

Box 3 | **Using archived specimens**

Advantages
- Convenience
- Speed
- Economy

Disadvantages
- Specimen quality can be affected by storage
- Patient information (e.g. age, sex and severity of symptoms) might be limited or not available
- Specific informed consent for such testing might not have been given at the time of specimen collection, so informed consent might need to be obtained or, if this is not possible, personal identifiers and patient information should be removed from specimens for testing

criteria might include recent use of antibiotics or other treatments. Such exclusions can make results easier to interpret but might also limit their ability to be applied generally to populations in which the test might be used in practice.

2.6. Sampling. The study group can consist of all subjects who satisfy the criteria for inclusion and are not disqualified by one or more of the exclusion criteria. In this case, a consecutive series of subjects is often included. Alternatively, the study group can be a sub-selection, for example, only those who test negative by the reference test. However, this can lead to biased estimates if the sample is not truly random.

2.7. Selecting the reference standard test. Where possible, all tests under evaluation should be compared with a reference (gold) standard. The choice of an appropriate reference standard is crucial for the legitimacy of the comparison. For example, a serological assay should not usually be compared with an assay that detects a microorganism directly, and clinically defined reference standards are not usually appropriate when clinical presentation is not sensitive or specific. Non-commercial or 'in-house' reference standards are legitimate only if they have been adequately validated. Sometimes, composite reference standards might have to be used in the absence of a single suitable reference standard. Results from two or more assays can be combined to produce a composite reference standard[6]. For example, if there are two possible 'gold standard' tests, both of which have high specificity but poorer sensitivity, then positives can be defined as samples that test positive by either test. In other circumstances, positives can be defined as those that test positive by both tests, negatives as those that test negative by both tests, and others omitted from the evaluation as indeterminate.

New tests under evaluation that are more sensitive than the existing reference standard usually require a composite reference standard. If a reference standard is not available and a composite standard cannot be constructed, an appropriate approach might be to report the levels of agreement between different tests, that is, positive by both or negative by both.

2.8. Evaluating more than one test. If more than one test is being evaluated, the evaluations can be sequential or simultaneous. The advantages and disadvantages of conducting simultaneous comparisons of several tests are listed in BOX 4.

2.9. Defining the outcome measures. The outcomes of the evaluation, such as performance characteristics, should be clearly defined. Evaluations should always include 95% confidence intervals for sensitivity and specificity (TABLE 1; Section 2.10).

In the absence of a reference standard, the performance of the test under evaluation can be compared to an existing test using a 2×2 table, which shows how the samples were classified by each test. The values for percentage agreement positive, percentage agreement negative, percentage negative by test 1 and positive by test 2, and percentage positive by test 1 and negative by test 2 can be derived from such a table. In addition, for prospective evaluations PPV and NPV can be used. These values will depend on the prevalence of the infection in the studied population.

In cases where the interpretation of test results is subjective, such as visual reading of a dipstick test result, an important outcome measure is assessment of the agreement between two or more independent readers.

2.10. Calculating the sample size. The key question to be addressed before embarking on a study, and the question that is often hardest to answer, is what level of performance is required of the test. The levels that might be acceptable in one setting might be inappropriate in another. The indications for performing the test can vary. The level and availability of healthcare resources and disease prevalence all have a bearing on setting the acceptable performance of a test.

Increasing the sample size reduces the uncertainty regarding the estimates of sensitivity and specificity (the extent of this uncertainty is summarized by the confidence interval). The narrower the confidence interval, the greater the precision of the estimate. A 95% confidence interval is often used — that is, we can be 95% certain that the interval contains the true values of sensitivity (or specificity). The formula for calculating the 95% confidence interval is given by equation 1

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}} \qquad (1)$$

where p = sensitivity (or specificity) measured as a proportion (not a percentage) and n = number of samples from infected people (or, for specificity, from uninfected people).

Box 4 | **Advantages and disadvantages of simultaneous comparisons**

Advantages
- Provide 'head to head' comparisons for two or more tests using a single reference standard and the same patient population
- Speed: results are available sooner than if conducting sequential trials
- More cost-effective

Disadvantages
- Can be difficult to interpret results for several tests independently, as blinding is not usually possible (that is, the results of other tests on the same samples or individuals might be known to the testers)
- Complicates the design of the evaluation, for example, randomization of the order in which specimens for the different tests are collected and assessed
- Specimen quality can be compromised with increasing numbers of specimens collected, especially with specimens collected from limited anatomical sites, such as urethral swabs or finger-prick blood specimens
- The collection of multiple specimens might not be acceptable to patients

Table 1 | **A 2 × 2 table to evaluate test performance**

Test under evaluation	Reference standard test		Total
	Positive	Negative	
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	

Test sensitivity = a/(a + c); test specificity = d/(b + d); PPV = a/(a + b); NPV = d/(c + d).
a = true positive, b = false positive; c = false negative; d = true negative

As an example of how confidence intervals are calculated, suppose 97 samples are positive by the 'gold standard' test and 90 of these are positive by the test under evaluation, then the sensitivity of the test is estimated by $p = 90/97 = 0.928$ and the confidence interval, using the formula above, is given in equation 2.

$$0.928 \pm 1.96 \times \sqrt{\frac{0.928(1 - 0.928)}{97}} = 0.928 \pm 0.051 = 0.877–0.979 \quad (2)$$

That is, we are 95% sure that the interval 87.7% to 97.9% contains the true sensitivity of the test under evaluation.

In considering sample size, it is important to consider the desired precision with which the sensitivity (or specificity) of the test is to be measured. To do this, we must first make a rough estimate of what we expect the sensitivity (or specificity) to be. So, if we suspect the sensitivity (or specificity) of the test under evaluation is approximately p (for example, 0.8 (80%)) and we wish to measure the sensitivity (or specificity) to within ± x (where x is expressed as a proportion rather than a percentage; for example, 0.10 rather than 10%) then we would choose n so that the confidence interval is ± x (for example ± 10%). This is shown in equations 3–5.

$$\sqrt{\frac{p(1 - p)}{n}} \leq x \quad (3)$$

which translates to:

$$n \geq \frac{(1.96)^2\, p(1 - p)}{x^2} \quad (4)$$

Thus, if $p = 0.80$ and $x = 0.10$, then

$$n \geq \frac{(1.96)^2\, 0.8(1 - 0.8)}{(0.1)^2} = 61.5 \quad (5)$$

Therefore, to measure the sensitivity to within ± 10% we require at least 62 samples that are positive by the 'gold standard' test.

TABLE 2 shows the relationship between sample size and 95% confidence interval for various estimated sensitivities and specificities. For example, if we estimate that the sensitivity of a new test is 80% and we want the confidence interval to be ± 6%, we will need to recruit, or have archived specimens from, 170 infected study subjects by the reference standard test. If the prevalence of infection in the study population is 10%, then there will be 10 infected subjects per 100 patients seen at the clinic. So, to have 170 infected subjects, we will need to recruit 1,700 patients (100/10 × 170).

In determining the sample size, allowance must also be made for patients who do not meet the inclusion criteria and the percentage who are likely to refuse to participate in the study.

If, when the study begins, it is not possible to estimate in advance what the sensitivity or specificity will be, then the safest option for the calculation of sample size is to assume these will be 50% (as this results in the largest sample size). Alternatively, sometimes it will be useful to conduct a pilot survey to estimate the prevalence of infection and to obtain a preliminary estimate of sensitivity and specificity. In such a study, the feasibility of the proposed study procedures can also be evaluated.

In some circumstances it might be possible to state the minimal acceptable sensitivity (or specificity) for the intended application of the test. So, if it is suspected that the sensitivity (or specificity) of the test under evaluation is p (for example, 80%) but it is considered that $p_0 = 70\%$ is the minimum acceptable sensitivity (or specificity), then n might be chosen so that the lower limit of the confidence interval is likely to exceed p_0. With the test requirement formulated in this way the sample size formula is given by equation 6:

$$n = (1.96 + 1.28)^2 \frac{p(1 - p)}{(p - p_0)^2} \quad (6)$$

For example, if it is anticipated that the sensitivity of a new test is 80% and to be acceptable for use in a given setting it must be at least 70%, then it will be necessary to recruit, or have archived specimens from, 168 infected study subjects. If the prevalence of infection in the study population is 10%, then it will be necessary to recruit a total sample of 1,680 (168/0.10), to ensure 168 infected individuals.

Details of methods for calculating sample size for diagnostic trials are available in REFS 5,7,8.

2.11. Evaluating reproducibility. The reproducibility of a test is a measure of the closeness of agreement between test results when the conditions for testing or measurement change. For example, reproducibility can be measured between operators (inter- and

Table 2 | **Relationship between sample size and 95% confidence interval**

Number of infected (non-infected) subjects required*	Estimated test sensitivity (or specificity)‡					
	50%	60%	70%	80%	90%	95%
50	13.9%	13.6%	12.7%	11.1%	8.3%	–
100	9.8%	9.6%	9.0%	7.8%	5.9%	4.3%
150	8.0%	7.8%	7.3%	6.4%	4.8%	3.5%
200	6.9%	6.8%	6.4%	5.5%	4.2%	3.0%
500	4.4%	4.3%	4.0%	3.5%	2.6%	1.9%
1,000	3.1%	3.0%	2.8%	2.5%	1.9%	1.4%

*As defined by the reference standard test. ‡95% confidence interval around the estimated sensitivity (+/– value in table).

Glossary

Accuracy
The percentage of correct results obtained by the test under evaluation compared with the results of a reference or 'gold standard' test. Usually expressed as the number of correct results divided by the total number of results, multiplied by 100.

Blinding
Interpreting a test result without knowledge of a patient's condition or previous test results.

Confidence interval
The confidence interval quantifies the uncertainty in measurement; usually reported as the 95% confidence interval, the range that we can be 95% certain covers the true value.

Negative predictive value (NPV)
The probability that a negative result accurately indicates the absence of infection.

Positive predictive value (PPV)
The probability that a positive result accurately indicates the presence of infection.

Prevalence
The proportion of a given population with an infection at a given time.

Proficiency panel
A collection of six or more mock or true specimens with positive and negative results for a particular test, used to ascertain the proficiency of the technologist in performing the test.

Quality assurance (QA)
An ongoing process of monitoring a system for reproducibility or reliability of results, with which corrective action can be instituted if standards are not met.

Reference standard
The best available approximation of a true result, generally indicating a test method that is currently accepted as reasonably, but not necessarily, 100% accurate. It is used as the reference method for assessing the performance characteristics of another test method.

Reproducibility
A measure of the extent to which replicate analyses using identical procedures agree with each other.

Sensitivity
The probability (percentage) that patients with the infection (determined by the result of the reference or 'gold standard' test) will have a positive result using the test under evaluation.

Specificity
The probability (percentage) that patients without the infection (determined by the result of the reference or 'gold standard' test) will have a negative result using the test under evaluation.

Tests
Any method for obtaining additional information regarding a patient's health status.

intra-observer reproducibility), between different test sites, using different instruments, between different kit lots (lot-to-lot reproducibility) or on different days (run-to-run and within-run reproducibility). The Kappa statistic[9] provides a useful measure of agreement between test types or lots, and between users. This statistic allows the measurement of agreement between sets of observations or test results above that expected by chance alone.

When test results are dichotomous (that is, either positive or negative), these characteristics are usually assessed in the following ways — operator-dependent reproducibility (especially important for tests for which the interpretation of results is subjective), in which the same lot of tests is assessed by two operators using the same evaluation panel but blinded to each other's results, and test-dependent reproducibility, which includes lot-to-lot variability, that is, the same operator evaluates different lots of diagnostic product using the same evaluation panel, and run-to-run variability, that is, the same operator evaluates the test several times using the same evaluation panel.

The repeatability of the test results refers to the closeness of the test results when no conditions of measurement change. The extent to which a test will produce the same result when used on the same specimen in identical circumstances (repeatability) should be distinguished from operator-related issues affecting reproducibility, which might be improved by further training.

The study protocol should describe how the reproducibility of the test will be measured and what aspect of reproducibility is being evaluated. This should include a description of the factors that are held constant, for example, reagent lots, instruments, calibration and/or quality-control methods. Reproducibility testing should be conducted in a blinded fashion, that is testers should not know the results obtained previously.

The size of the evaluation panel for reproducibility studies should be dictated by the degree of precision needed for the relevant clinical indication. The panel should include at least one positive and one negative control, and, if appropriate, two or three different operators, with the samples evaluated on three different days. In multi-centre studies, reproducibility should be assessed at each centre and between centres.

As well as measuring the extent to which there is reproducibility in the assessment of strong positive results, it is important to include assessment of reproducibility using weak positive and borderline negative samples if these might be important for clinical decision making.

2.12. Evaluating operational characteristics. An evaluation of a test can also include an assessment of its operational characteristics and cost-effectiveness. The latter is not considered in this document. Some operational characteristics, such as simplicity, acceptability of the test to users, the robustness of the test under different storage conditions and the clarity of instructions, are qualitative and subjective, but assessment of these can be crucial for decisions regarding the suitability of a test for a specific setting. In particular, the robustness of the test under different storage conditions is an area of concern for tests that will be used in remote settings.

Diagnostic tests can contain biological or chemical reagents that are heat-labile and might be affected by moisture, making the shelf-life of the test dependent on the temperature and humidity at which it is stored. Many commercially available *in vitro* diagnostic tests are recommended to be stored between 4°C and 30°C and are sealed in moisture-proof packaging. The specified shelf-life is based on the assumption that these conditions are maintained. Transport and operational conditions in the tropics commonly exceed 30°C, especially for point-of-care tests used in remote areas. Exposure to humidity can occur during delays between opening of the moisture-proof packaging and performance of the test procedure.

During evaluation of diagnostic tests, it is essential to inspect test kits for signs of damage caused by heat or humidity, and to record the conditions under which the tests have been stored and transported. These conditions should be taken into account when interpreting the results. A product dossier of test characteristics, including heat-stability data, should be available from the manufacturer of the diagnostic test. This will assist in extrapolating the results obtained under trial conditions to the results expected if the test kits had been stored under the anticipated operational conditions.

If there is uncertainty about the test stability, storage outside the manufacturer's recommendations is expected during operational use or there are insufficient data on temperature stability, the addition of thermal-stability testing to the trial protocol should be considered. Tests can be stored in an incubator at temperatures near the likely maximum in the field (for example, 40°C for 2–3 months), then assessed in comparison

Box 5 | **Designing a protocol for an evaluation using archived specimens**

1. Define the target population for the test under evaluation
2. Define the type of specimens that should be included in the evaluation panel
3. Define how appropriate specimens should be selected for the evaluation panel and how the specimens should have been stored
4. Calculate the required sample size
5. Develop a method to remove personal identifiers from the specimens (unless previous consent has been given for this type of work) by assigning a study code to each specimen
6. Define how the specimens will be tested to ensure that the results of the reference standard test will not be known when performing the test under evaluation, and vice versa ('blinding')
7. Define a plan to ensure proficiency in performing the reference standard test
8. Define a plan for quality assurance and external validation of trial results
9. Define where the study protocol needs to be sent for ethics approval (local and other relevant ethics committees)
10. Develop a data-analysis plan, for the calculation of sensitivity, specificity and confidence intervals
11. Define the methods for the dissemination of trial results

with tests stored at the recommended temperature during this period. During field evaluations, periodic comparison of the performance of tests stored at ambient temperature in the field against those stored at recommended temperatures should give an indication of the thermal stability of the test and it might be appropriate to stop the evaluation if the results show substantial deterioration of tests.

2.13. Quality assurance and monitoring. All studies should incorporate quality assurance (QA). Study QA procedures should be established to ensure that the studies are conducted and the data are generated, documented and reported in compliance with good clinical laboratory practice (GCLP). GCLP, rather than good laboratory practice (GLP), is more appropriate for trials that are not being undertaken for registration (see http://www.qualogy.co.uk) or for applicable regulatory requirement purposes. QA should be overseen by an individual who is not a member of the study team.

In the context of an evaluation trial, QA comprises:

- Study quality control (SQC): the crucial element of SQC is the generation of, and adherence to, standard operating procedures (SOPs), which comprise detailed and specific written instructions as to how all aspects of the study are to be conducted[10].
- External quality monitoring (EQM): independent monitoring of quality, which can include site visits conducted by a trained study monitor from outside the study team.
- Study quality improvement (SQI): the process through which deficiencies identified through SQC and EQM are remedied.

QA of laboratory and/or diagnostic testing procedures is also crucial in the day-to-day running of a diagnostic laboratory. Laboratory QA comprises internal quality control (IQC), external quality assessment (EQA) and quality improvement measures. IQC refers to the internal measures taken to ensure that laboratory results are reliable and correct, for example, the existence of SOPs for each test procedure, positive and negative controls for assays, stock management to prevent expired reagents being used, and monitoring of specimen quality. EQA, which is sometimes referred to as proficiency testing, is an external assessment of the laboratory's ability to maintain satisfactory quality, ensured by regular testing of an externally generated panel of specimens. Quality improvement is the process through which deficiencies identified through IQC or EQA are remedied and includes staff-training sessions, recalibration of equipment and assessment of the positive and negative controls used for particular tests.

IV. THE DESIGN OF DIAGNOSTIC EVALUATIONS USING ARCHIVED SPECIMENS

If the test evaluation can be undertaken satisfactorily using archived specimens and a panel of well-characterized specimens is available, a retrospective evaluation can be conducted with both the new test and the reference standard. Although this type of study has the advantages of being rapid and relatively inexpensive compared with a prospective study, it is important to consider several factors that might limit the generalizability of the results, including whether the specimens were collected from a population similar to the population in which the test will be used; what clinical and laboratory results are available to characterize the specimens; whether the specimens have been stored appropriately; and whether there are sufficient numbers of positive and negative specimens to provide an adequate sample size.

The steps involved in designing a protocol for an evaluation using archived specimens are outlined in BOX 5.

Details of this information and the procedures to be followed should be stated in the study protocol. External validation can be performed by sending all positive specimens and a proportion of the negative specimens to another laboratory for testing. Informed consent is usually not required for trials using archived specimens from which personal identifiers have been removed. Some ethics review committees require the investigator to provide information on how the specimens can be

Box 6 | **Designing a protocol for a prospective evaluation**

1. Define the target population for the test under evaluation
2. Develop methods for the recruitment of study participants and informed consent procedures
3. Design study instruments such as data-collection forms and questionnaires
4. Develop plans to pilot study instruments to determine whether they are appropriate
5. Calculate the required sample size
6. Develop a plan for specimen collection, handling, transport and storage
7. Define how the specimens will be tested to ensure blinding of results of the reference standard test from the results of the test under evaluation
8. Define a plan to ensure proficiency in performing the reference standard test
9. Develop a data-collection and data-analysis plan
10. Develop plans to ensure the confidentiality of study data
11. Define a plan for quality assurance and external validation of trial results
12. Define where the study protocol needs to be sent for ethics approval (local and other relevant ethics committees)
13. Define methods for the dissemination of trial results

Box 7 | **Information to be included in the Patient Information and Consent Forms**

- Purpose of the study
- Study procedures and what is required of participants
- Assurance that participation is voluntary
- Statement of the possible discomfort and risks of participation
- Benefits (for example, treatment or care to be offered to those who test positive by the reference standard test)
- Compensation offered for travel and other out-of-pocket expenses
- Safeguards to ensure confidentiality of patient information
- Freedom to refuse to participate and alternatives to participation, and freedom to withdraw from the study at any time without compromise to future care at the facility
- Use of study data and publication of results
- Contact details of a locally accessible person who can answer questions from participants for the duration of the study
- Participant statement to indicate that they understand what was explained to them and they agree to participate by signing the consent form. Illiterate participants can give consent by a thumbprint witnessed by a third party

made anonymous and require assurance that results cannot be traced to individual patients.

V. THE DESIGN OF PROSPECTIVE DIAGNOSTIC EVALUATIONS

The recommended steps in designing a prospective diagnostic evaluation are outlined in BOX 6.

1. Defining the target population for the test under evaluation

The characteristics of the study population should be fully described (see section III, 2.1)

2. Developing methods for the recruitment of study participants and informed consent procedures

Consider the following:

- Who recruits the study subjects? Ideally, this should not be the clinician caring for the participants, as this might influence the participants' decision.
- Who is eligible for enrolment?
- How will informed consent be obtained? (Recruitment of children will require approval from a parent or guardian.)
- Who will answer participants' questions about the study?
- How will confidentiality be assured?

Further information on informed consent can be obtained from REFS 11 & 12.

The Patient Information and Consent Forms should be clear, concise and in a language (read or narrated) that is understandable to the patient. The forms should include the points outlined in BOX 7. An example consent form is shown in APPENDIX 2. Templates are also available from many academic research ethics review committee websites including the WHO Research Ethics Review Committee (http://www.who.int/rpc/research_ethics/en/). If biological specimens are to be stored for future use, this should be specified in a separate section in the consent form and participants should be given the option to refuse to have their specimens stored but still participate in the study.

In general, the only payment to study subjects should be for compensation for transport to the clinic and loss of earnings because of clinic visits related to the study. Payment should never be excessive, such that it might constitute an undue incentive to participate in the study.

Treatment should usually be provided free of charge. Any treatment decisions (if appropriate) should not be based on the results of the test under evaluation but on the reference test. Refusal to participate in the study should not prejudice access to treatment that would normally be accessible.

3. Designing study instruments

Each item on the patient data form should be considered with respect to the stated aims and objectives of the trial. The collection of unnecessary data is a waste of resources and might detract attention from recording the most important data.

When designing data-record forms, it is advisable to review forms from similar trials; allow adequate time to design, translate (and back-translate) and pilot data forms before starting data collection; specify who will complete each form (interviewer or study subject); and specify the QA procedures to ensure data are recorded correctly.

The layout and content of forms should be discussed with the study staff who will be responsible for data management and analysis. The forms should be user-friendly and data should be easily entered into a database. Consider the paper size and colour, the format of records (record books with numbered pages are preferable to loose sheets of paper) and the use of boxes or lines for multiple-choice responses. Questions can allow open or structured responses. Structured responses limit allowable responses to a predefined list, whereas open responses allow freedom to record unanticipated answers, but are more difficult to code and analyse.

It should be ensured that those who will be completing the forms fully understand the forms and know how to complete the forms correctly. Clarity of language is important, particularly when translation might be necessary and so the forms should use simple, uncomplicated language; avoid abbreviations, ambiguous terms and acronyms; avoid unnecessary wording and compound questions; provide definitions; and translate (and back-translate) all of the questions to ensure the correct data items are recorded.

Ensure that a distinction can be made between omitted responses and responses such as 'not applicable'. Where items are to be skipped, the form should contain documentation of the legitimacy of a skipped answer.

4. Develop plans to pilot study instruments

Plans should be developed to determine whether the study instruments, such as questionnaires and data-collection forms, are appropriate. Questions might need to be rephrased to obtain the relevant response. So far as is possible, all aspects of the study should be piloted in a small study so that the methods and procedures can be modified as appropriate. The pilot study also provides the ability to make a preliminary estimate of infection prevalence, which might aid planning the size of the main study.

5. Calculating the required sample size

See Section III, 2.10.

6. Developing a plan for the study logistics

A plan should be developed for safe specimen collection, handling, transport and storage. Consider using pre-printed unique study numbers for forms and specimens (labels should be tested for adherence when samples are frozen, if necessary). Also, develop a flow diagram for specimen handling that can be distributed to laboratory staff.

7. Defining the blinding of results
Specimens will be tested to ensure blinding of results of the reference standard test from the results of the test under evaluation. Most rapid tests require subjective interpretation of the test result. Steps must be taken to ensure that the staff performing the reference test are not aware of the results from the test under evaluation, and vice versa. Also, laboratory staff should not be aware of clinical findings or of the results of other laboratory tests.

It can be difficult to ensure blinding if several tests are being evaluated at the same time. For any repetitively performed procedures, consider randomizing the order in which they are done —for example, if multiple swabs are to be taken, consider applying the tests in random order to different swabs.

8. Defining a QA plan
A QA plan should be developed for quality management of the diagnostic trial. This includes ensuring that the study personnel are proficient in performing both the tests under evaluation and the reference standard test. Before the start of the trial, the laboratory (or whoever is to perform the tests) should be able to demonstrate proficiency in performing the reference standard test(s). The personnel performing the test under evaluation should also demonstrate proficiency at performing and reading this test. The laboratory should subscribe to external proficiency programmes where available. Training records of study personnel should be kept. The QA plan should also include quality management of study data.

9. Developing a plan for data collection and data analysis
Study results entered into workbooks or directly into computer spreadsheets should be checked daily and signed off by the clinic or laboratory supervisor if possible. When entering results into a computer database, consider double data entry to minimize inadvertent errors. All records and study data should be backed up regularly, preferably daily. Review processes for the study database and approval mechanisms for items to be added or deleted should be established. The form of the tables that will be used in the analysis and the statistical methods that will be used in the interpretation of the study results should be drafted before data have been collected to ensure that all the relevant information will be recorded.

10. Developing plans to ensure the confidentiality of study data
All study data should be kept confidential (for example, in a locked cabinet and a password-protected database, with access limited to designated study personnel).

11. Defining a plan for external validation of trial results
See Section III, 2.13.

12. Scientific and ethical review of study protocol
The study protocol should undergo scientific and ethical review by the relevant bodies. Submission documents for ethics approval must follow national or institutional guidelines. As a minimum, the application document for ethics committee approval should contain the information shown in BOX 8. In addition, some ethics committees require protocols to have undergone prior scientific review.

13. Defining methods for the dissemination of trial results
This can involve submitting results for publication to a scientific journal, but most importantly, there should be a plan to inform those responsible for procuring or authorizing tests of the study findings. Appropriate feedback should be given to study participants.

Box 8 | **Information required in the application document for ethics committees**

- Statement of study objectives and rationale
- Description of study methods
- Preliminary evidence of safety and efficacy
- Type/source of patients or samples
- Primary outcome measure
- Follow up of patients
- Sample size plus rationale for proposed size
- Randomization and method of assignment, if applicable
- Risks and benefits for those participating in the study
- Methods to protect patients from harm
- Safeguards for patient privacy and confidentiality
- Benefits expected to be derived from the study
- Alternatives to participation
- Contact details of a locally accessible person who can answer questions from participants for the duration of the study
- Dissemination of study results and any other relevant material

VI. SITE SELECTION AND STUDY PREPARATION

1. Criteria for selection of field sites
The criteria for field-site selection can include:
- Easy access to suitable target populations.
- Adequate prevalence of infection/disease so that sufficient numbers of infected (and uninfected) people can be recruited.
- Availability of suitably trained study personnel (sometimes further training might be required for the purposes of the trial).
- Adequate facilities for conducting the study, for example, space for conducting confidential interviews.
- Good standard of care available for people found to be infected.
- Capacity to store specimens in correct conditions.
- Sufficient data-handling capacity (for example, staff and computers).
- Ability to perform data analysis (on site, if possible).
- Access to good laboratory facilities (if relevant laboratory accreditation schemes exist, and the laboratory is eligible, it should be accredited).
- A mechanism for ethical review and approval of the trial protocol.

2. Site preparation
2.1. Setting up a trial-management system. From the outset of the trial, a quality-management system should be in place. The composition of the trial team should be clearly defined, as should the responsibilities of each team member and trial-management and trial-monitoring procedures.

2.2. Preparing SOPs. SOPs should be prepared for for all clinical and laboratory procedures required in the study protocol (see REF. 10 and BOX 9).

2.3. Training workshops for GCP and GLP/GCLP. Before the trial begins, the study team should be given training on the principles and implementation of GCP and GLP/GCLP[13].

Box 9 | **Elements to be included in SOPs**

- Recruitment of study participants
- Specimen collection, handling, storage and transport
- Preparation of reagents
- How to use test kits and interpret test results, including handling of indeterminate results
- How to perform reference standard tests
- How to monitor and calibrate equipment
- How to identify and correct malfunctions or errors
- Specific instructions on quality assurance procedures
- Record keeping of trial results

2.4. Assurance of proficiency at performing reference standard and tests under evaluation. Before the trial starts, the laboratory should be able to demonstrate proficiency in performing the reference standard tests as well as the tests under evaluation. The laboratory should subscribe to external proficiency programmes. Training records of study personnel should be kept. Training should be provided for performing the test under evaluation using well-characterized positive and negative specimens.

2.5. Piloting and refining study instruments, including the informed consent process. This is essential to ensure the information is understood by study participants and that the questions are appropriate. Translation of the informed consent information sheet into the local language is also essential. Back-translation is desirable to ensure the accuracy of the information provided to the study participants to allow them to make an informed decision whether or not to participate in the study.

VII. CONDUCTING THE EVALUATION

1. General guidelines on the use of test kits

- Note the lot number and expiry date; a kit should not be used beyond the expiry date.
- Ensure correct storage conditions are in place, as stated by the manufacturer. If this is not possible in the field, or cannot be ensured during transport, this should be made clear when the study is reported. If a desiccant is included in the package, the kit should not be used if the desiccant has changed colour.
- Generally, if test kits are stored in a refrigerator, they should be brought to room temperature approximately 30 minutes before use. The use of cold test kits can lead to false-negative results.
- Damaged kits should be discarded.
- Open test kits only when they have reached room temperature, unless otherwise specified.
- Use test kits immediately after opening.
- Reagents from one kit should not be used with those from another kit.
- Tests should be performed exactly as described in the product insert (if available) or any variations must be clearly noted, such as the method of transferring the sample to the kit or the use of venous blood rather than a finger-prick sample.
- It can be useful to evaluate 'off-label' use: this refers to the use of a test for an indication or with a specimen not mentioned in the package insert, for example, self-administered vaginal swabs or pharyngeal swabs. This can be important in defined circumstances, but the fact that it is off-label use must be clearly stated when the results are reported.

2. Biosafety issues

The investigators must comply with national workplace safety guidelines with regard to the safety of clinic and laboratory personnel and the disposal of infectious waste. General guidelines are given in BOX 10.

3. Trial management

3.1. The facility and equipment. Laboratory facilities and equipment should be available and adequately maintained for the work required, for example, suitable work areas, lighting, storage, ventilation and hand-washing facilities should be available. Where field conditions necessitate different standards of operation, these should be clearly stated in the protocol.

3.2. Proficiency of personnel. There are various options for external QA or proficiency programmes for certain infectious diseases such as the College of American Pathologists Inter-laboratory Survey Programs (http://www.cap.org/apps/cap.portal) or the United Kingdom National External Quality Assessment Service (http://www.ukneqas.org.uk). Ongoing records of performance of proficiency panels should be kept to monitor proficiency, especially when there is a change of personnel.

3.3. Changes of study procedures. Any changes to study procedures should be accompanied by changes in the relevant SOPs. Changes to SOPs should be documented, signed off by the responsible supervisor, dated and disseminated to the study team.

4. Quality assurance

There should be arrangements in place (a QA unit or designated person) to ensure that the study is conducted in accordance with the study protocol. A system should be established so that corrective actions suggested to the study team are properly and rapidly implemented.

5. Trial monitoring

There should be regular independent assessment of the laboratory and/or field team performing the evaluations in compliance with the principles of GCP and GLP/GCLP, including both internal and external quality control and QA procedures.

6. Data analysis

The data should be analysed according to the analysis plan after checking, and if necessary correcting, the study data. The sensitivity and specificity of a test can be calculated by comparing the test results to the validated reference test results. They can be displayed in a 2×2 table, as illustrated in TABLE 1. In addition, for prospective trials, the PPV (a/(a + b)) and NPV (d/(c + d)) can be calculated. Inter-observer variability is calculated as the number of tests for which different results are obtained by two independent readers, divided by the number of specimens tested.

Box 10 | **General biosafety guidelines**

- Treat all specimens as potentially infectious
- Wear protective gloves and a laboratory gown while handling specimens
- Do not eat, drink or smoke in the laboratory
- Do not wear open-toe footwear in the laboratory
- Clean up spills with appropriate disinfectants
- Decontaminate all materials with an appropriate disinfectant
- Dispose of all waste, including all clinical material and test kits, using an appropriate method such as placing sharp objects in a biohazard container and disposable materials in sealable waste bags for incineration

VIII. REPORTING AND DISSEMINATING RESULTS

Wherever possible, study participants should be given feedback on study results by, for example, meeting with the study community or having a readily accessible contact person at a clinic to answer specific queries. The results can also be disseminated by publication in peer-reviewed journals or posted on relevant websites. The STARD checklist should be used to guide how a study is reported (APPENDIX 1).

Currently, published studies vary in their attainment of the STARD criteria, often succumbing to common pitfalls[14–16] including inadequate data being used as evidence (including inadequate sample size); bias (for example, by poor selection of study subjects, inappropriate representation of the intended target population, lack of blinding or the use of poor or no reference standards); inadequate description of the characteristics of the study population (for example, parasite density can affect the sensitivity of malaria tests); and evaluations in populations for which the tests are not intended.

IX. CONCLUSIONS

The rapid advances that have been made in molecular biology and molecular methods have led, and continue to lead, to the development of sensitive and specific diagnostic tests, which hold the promise of substantially strengthening our ability to diagnose, treat and control many of the major infectious diseases in developing countries. It is imperative that these new diagnostics are rigorously and properly evaluated in the situations in which they will be deployed in disease control before they are released for general use. A poorly performing diagnostic might not only waste resources but might also impede disease control. The basic procedures described in this article for designing and conducting diagnostic evaluations provide an outline for ensuring the proper evaluation of new diagnostics in laboratory and field trials.

Shabir Banoo is at the Medicines Control Council of South Africa, Pretoria, South Africa.

David Bell is at the Malaria and other Vector-borne and Parasitic Diseases, World Health Organization–Regional Office for the Western Pacific, Manila, Philippines.

Patrick Bossuyt is at the Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands.

Alan Herring is at the Veterinary School, University of Bristol, Bristol, UK.

David Mabey is at the Clinical Research Unit, London School of Hygiene and Tropical Medicine, London, UK.

Freddie Poole is at the Division of Microbiology Devices, Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, USA.

Peter G. Smith is at the Infectious Diseases Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK.

N. Sriram is at the Tulip Group of Companies, Goa, India.

Chansuda Wongsrichanalai is at the US Naval Medical Research Unit 2, Jakarta, Indonesia.

Ralf Linke, Rick O'Brien and Mark Perkins are all at the Foundation for Innovative Diagnostics (FIND), Geneva, Switzerland.

Jane Cunningham, Precious Matsoso, Carl Michael Nathanson, Piero Olliaro, Rosanna W. Peeling and Andy Ramsay are all at the UNICEF/UNDP/World Bank/WHO Special Programme for Research & Training in Tropical Diseases (TDR), World Health Organization, Geneva, Switzerland.*

**e-mail: peelingr@who.int*

doi: 10.1038/nrmicro1523

1. Greenhalgh, T. How to read a paper: papers that report diagnostic or screening tests. *Br. Med. J.* **315**, 540–543 (1997).
2. Borriello, S. P. Near-patient microbiological tests. *Br. Med. J.* **319**, 298–301 (1999).
3. Bossuyt, P. M. *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Clin. Chem.* **49**, 1–6 (2003).
4. Bossuyt, P. M. *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin. Chem.* **49**, 7–18 (2003).
5. Smith, P. G. & Morrow, R. H., eds *Field Trials of Health Interventions in Developing Countries: A Toolbox*, (Macmillan, London, 1996).
6. Alonzo, T. D & Pepe, M. S. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statist. Med.* **18**, 2987–3003 (1999).
7. Pepe, M. S. *Statistical Evaluation of Medical Tests for Classification and Prediction*. (Oxford Univ. Press, 2003).
8. Gardner, M. J. & Altman, D. G. Estimating with confidence. *Br. Med. J.* **296**, 1210–1211 (1988).
9. McGinn, T. *et al.* Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *Can. Med. Assoc. J.* **171**, 1369–1373 (2004).
10. WHO. *Standard Operating Procedures for Clinical Investigators*. UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (WHO/TDR). TDR/ TDP/SOP99.1 [online] < http://www.who.int/tdr/publications/publications/pdf/sop.pdf > (1999)
11. WHO/Council for International Organizations of Medical Sciences. *International Ethical Guidelines for Biomedical Research Involving Human Subjects* (2002).
12. Nuffield Council on Bioethics. *The Ethics of Research Related to Healthcare in Developing Countries* [online] < http://www.nuffieldbioethics.org/go/ourwork/developingcountries/publication_309.html > (2002)
13. WHO. *Guidelines for Good Laboratory Practice*. UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases TDR/ PRD/GLP/01.2 (WHO/TDR,Geneva, 2001).
14. Delaney, B. C. *et al*. Systematic review of near-patient test evaluations in primary care. *Br. Med. J.* **319**, 824–827 (1999).
15. Reid, M. C., Lachs, M. S. & Feinstein, A. Use of methodological standards in diagnostic test research. Getting better but still not good. *J. Amer. Med. Assoc.* **274**, 645–651 (1995).
16. Small, P. M. & Perkins, M. D. More rigour needed in trials of new diagnostic agents for tuberculosis. *Lancet* **356**, 1048–1049 (2000).

Acknowledgements

We wish to thank Izabela Suder-Dayao for excellent secretarial support, and Robert Ridley and Giorgio Roscigno for support and guidance.

APPENDIX 1 | STANDARDS FOR REPORTING OF DIAGNOSTIC ACCURACY (STARD) CHECKLIST

Section and topic	Item #		On page #
Title/abstract/ keywords	1	Identify the article as a study of diagnostic accuracy (recommended MeSH heading 'sensitivity and specificity').	❑
Introduction	2	State the research questions or study aims, such as estimating the diagnostic accuracy or comparing accuracy between tests or across participant groups.	❑
Methods		Describe:	
Participants	3	The study population: the inclusion and exclusion criteria, the setting and the locations where the data were collected.	❑
	4	Participant recruitment: was the recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	❑
	5	Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.	❑
	6	Data collection: was data collection planned before (prospective study) or after (retrospective study) the index test and reference standard were performed?	❑
Test methods	7	The reference standard and its rationale.	❑
	8	Technical specifications of the material and methods involved, including how and when the measurements were taken, and/or cite references for the index tests and reference standard.	❑
	9	Definition of, and rationale for, the units, cut offs and/or categories of the results of the index tests and the reference standard.	❑
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	❑
	11	Whether or not the readers of the index tests and reference standard were blind to the results of the other test and describe any other clinical information available to the readers.	❑
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	❑
	13	Methods for calculating test reproducibility, if done.	❑
Results		Report:	❑
Participants	14	When the study was done, including the start and end dates of recruitment.	❑
	15	The clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co-morbidity, current treatments and recruitment centres).	❑
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	❑
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered inbetween.	❑
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	❑
	19*	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard.	❑
	20	Any adverse events from performing the index tests or the reference standard.	❑
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	❑
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	❑
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.	❑
	24	Estimates of test reproducibility, if done.	❑
	25	Discuss the clinical applicability of the study findings.	❑

* This entry has been modified from the original.

APPENDIX 2 | SAMPLE INFORMED CONSENT FORM

(A separate patient information sheet containing this information should also be provided)

A | PURPOSE OF THE STUDY

Chlamydial infection is caused by bacteria that are transmitted by sexual intercourse. In women, this infection can cause pelvic pain and, in the long term, increase the risk of infertility. Furthermore, during unprotected sexual intercourse with a man infected with the AIDS virus, a woman infected with chlamydia will have a higher risk of acquiring the AIDS virus than a woman not infected with this bacterium.

To find out whether you have this infection, we need to do some laboratory tests. These tests are expensive and the results are not available the same day. Rapid tests to diagnose chlamydia within 30 minutes are now available but we do not know if they are accurate or reliable. The main purpose of this study is to evaluate a rapid test for the diagnosis of chlamydial infection. We would like to compare the result of these rapid tests with a laboratory-based test to see if they are as accurate as laboratory tests.

B | STUDY PROCEDURES

If you agree to participate in the study, you will be assigned a study number. The doctor or nurse will give you a physical examination and ask you some questions according to standard clinic procedure. He/she will take two samples from your vagina and two samples from your cervix. Your name will not appear on any samples or on the questionnaire. All the samples will be destroyed at the end of the study. If you are diagnosed with chlamydial infection using the standard laboratory tests, you will be treated with antibiotics on your follow-up visit according to normal clinic procedure. You will not be treated according to the rapid test results as we are not yet sure if it is accurate.

C | VOLUNTARY PARTICIPATION

Your decision not to participate in this study will not affect the care you will receive at the clinic in any way. Even if you do agree to become a study participant, you can withdraw from the study at any time (verbally) without affecting the care that you will receive. During the interview, you can choose not to answer any particular question.

D | DISCOMFORT AND RISKS

You may feel a small amount of discomfort or have a small amount of bleeding from the vagina after the pelvic examination and specimen collection.

E | BENEFITS

There will be no immediate benefits from your participation in the study. When the study results are known and if the rapid tests are acceptable in terms of accuracy, everyone who comes to the clinic could benefit from having this test available to diagnose chlamydia and receive the right treatment the same day.

F | COMPENSATION

There will be no monetary compensation for this study, but routine medical consultation and appropriate referral services are available.

G | CONFIDENTIALITY STATEMENT

The records concerning your participation are to be used only for the purpose of this research project. Your name will not be used on any study forms or labels on laboratory specimens or in any report resulting from this study. At the beginning of the study, we will give you a study identification number and this number will be used on the forms and on the laboratory specimens. Any information obtained in connection with this study will be kept strictly confidential. Only members of the study team will have access to information linking your name with your study number.

H | QUESTIONS AND FREEDOM TO WITHDRAW FROM THE STUDY

You can withdraw from the study at any time without affecting your present or future medical care at the clinic. You can contact any of the study personnel if you have questions about the research. (Please give the contact name, address and telephone number of the contact person for each site).

I | RESULTS PUBLICATION

When the researchers have analysed the data, the results and the explanation of its implications will be posted at the clinic for everyone's information.

J | PARTICIPANT STATEMENT

I have been informed verbally and in writing about this study and understand what is involved. I also know whom to contact if I need more information. I understand that confidentiality will be preserved. I understand that I am free to withdraw from the study at any time without affecting the care I normally receive at the clinic. I agree to participate in this study as a volunteer subject and will be given a copy of this information sheet to keep.

_____ _____

Date Name of participant

_____

Signature (or thumb print or cross) of participant

_____ _____

Date Name of witness

_____

Signature of witness

K | INVESTIGATOR'S STATEMENT

I, the undersigned, have defined and explained to the volunteer in a language she understands, the procedures of this study, its aims and the risks and benefits associated with her participation. I have informed the volunteer that confidentiality will be preserved, that she is free to withdraw from the trial at any time without affecting the care she will receive at the clinic. Following my definitions and explanations the volunteer agrees to participate in this study.

_____ _____

Date Name of investigator who gave the information about the study

Signature: _____

Sources of Variation and Bias in Studies of Diagnostic Accuracy

A Systematic Review

Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Afina S. Glas, MD, PhD; Patrick M.M. Bossuyt, PhD; and Jos Kleijnen, MD, PhD

Background: **Studies of diagnostic accuracy are subject to different sources of bias and variation than studies that evaluate the effectiveness of an intervention. Little is known about the effects of these sources of bias and variation.**

Purpose: **To summarize the evidence on factors that can lead to bias or variation in the results of diagnostic accuracy studies.**

Data Sources: **MEDLINE, EMBASE, and BIOSIS, and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration. Methodologic experts in diagnostic tests were contacted.**

Study Selection: **Studies that investigated the effects of bias and variation on measures of test performance were eligible for inclusion, which was assessed by one reviewer and checked by a second reviewer. Discrepancies were resolved through discussion.**

Data Extraction: **Data extraction was conducted by one reviewer and checked by a second reviewer.**

Data Synthesis: **The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias, the amount of evidence was limited. Evidence was lacking for other features, including incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts.**

Conclusions: **Many issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation; however, the empirical evidence about the size and effect of these issues is limited.**

Ann Intern Med. 2004;140:189-202. **www.annals.org**
For author affiliations, see end of text.

Diagnostic tests are of crucial importance in health care. They are performed to reduce uncertainty concerning whether a patient has a condition of interest. A thorough evaluation of diagnostic tests is necessary to ensure that only accurate tests are used in practice. Diagnostic accuracy studies are a vital step in this evaluation process.

Diagnostic accuracy studies aim to investigate how well the results from a test being evaluated (index test) agree with the results of the reference standard. The reference standard is considered the best available method to establish the presence or absence of a condition (target condition). In a classic diagnostic accuracy study, a consecutive series of patients who are suspected of having the target condition undergo the index test; then, all patients are verified by the same reference standard. The index test and reference standard are then read by persons blinded to the results of each, and various measures of agreement are calculated (for example, sensitivity, specificity, likelihood ratios, and diagnostic odds ratios).

This classic design has many variations, including differences in the way patients are selected for the study, in test protocol, in the verification of patients, and in the way the index test and reference standard are read. Some of these differences may bias the results of a study, whereas others may limit the applicability of results. Bias is said to be present in a study if distortion is introduced as a consequence of defects in the design or conduct of a study. Therefore, a biased diagnostic accuracy study will produce estimates of test performance that differ from the true performance of the test. In contrast, variability arises from differences among studies, for example, in terms of population, setting, test protocol, or definition of the target disorder (1). Although variability does not lead to biased estimates of test performance, it may limit the applicability of results and thus is an important consideration when evaluating studies of diagnostic accuracy.

The distinction between bias and variation is not always straightforward, and the use of different definitions in the literature further complicates this issue. For example, when a diagnostic study starts by including patients who have already received a diagnosis of the target condition and uses a group of healthy volunteers as the control group, it is likely that both sensitivity and specificity will be higher than they would be in a study made up of patients only suspected of having the target condition. This feature has been described as spectrum bias. However, strictly speaking, one could argue that it is a form of variability; sensitivity and specificity have been measured correctly within the study and thus there is no bias; however, the results cannot be applied to the clinical setting. In other words, they lack generalizability (2). Others have argued that when the goal of a study is to measure the accuracy of a test in the clinical setting, an error in the method of patient selection is made that will lead to biased estimates of test performance. They use a broader definition of bias and take into account the underlying research question when deciding whether results are biased. In this paper, we use a more restricted definition of bias.

Our goal is to classify the various sources of variation and bias, describe their effects on test results, and provide a summary of the available evidence that supports each source of bias and variation (**Table 1**). For this purpose, we

Table 1. **Description of Sources of Bias and Variation**

Source	Bias or Variation	Description
Population		
Demographic features	Variation	Tests may perform differently in various samples. Therefore, demographic features may lead to variations in estimates of test performance.
Disease severity	Variation	Differences in disease severity among studies may lead to differences in estimates of test performance.
Disease prevalence	Variation	The prevalence of the target condition varies according to setting and may affect estimates of test performance. Context bias, the tendency of interpreters to consider test results to be positive more frequently in settings with higher disease prevalence, may also affect estimates of test performance.
Distorted selection of participants	Variation	The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice, the results of the study may have limited applicability.
Test protocol: materials and methods		
Test execution	Variation	A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy can be the result of differences in test execution.
Test technology	Variation	When the characteristics of a diagnostic test change over time as a result of technological improvement or the experience of the operator of the test, estimates of test performance may be affected.
Treatment paradox and disease progression bias	Bias	Disease progression bias occurs when the index test is performed an unusually long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed. Treatment paradox occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started.
Reference standard and verification procedure		
Inappropriate reference standard	Bias	Errors of imperfect reference standard or standards bias the measurement of diagnostic accuracy of the index test.
Differential verification bias	Bias	Part of the index test results is verified by a different reference standard.
Partial verification bias	Bias	Only a selected sample of patients who underwent the index test is verified by the reference standard.
Interpretation (reading process)		
Review bias	Bias	Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.
Clinical review bias	Bias	The availability of information on clinical data, such as age, sex, and symptoms, during interpretation of test results may affect estimates of test performance.
Incorporation bias	Bias	The result of the index test is used to establish the final diagnosis.
Observer variability	Variation	The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. In 2 or more observations of the same diagnostic study, intraobserver variability occurs when the same person obtains different results, and interobserver variability occurs when 2 or more people disagree.
Analysis		
Handling of indeterminate results	Bias	A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.
Arbitrary choice of threshold value	Variation	The selection of the threshold value for the index test that maximizes the sensitivity and specificity of the test may lead to overoptimistic measures of test performance. The performance of this cutoff in an independent set of patients may not be the same as in the original study.

conducted a systematic review of all studies in which the main focus was examine the effects of one or more sources of bias or variation on estimates of test performance.

Methods

Literature Searches

We searched MEDLINE, EMBASE, BIOSIS and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration from database inception to 2001. Search terms included *sensitivit**, *mass-screening*, *diagnostic-test*, *laboratory-diagnosis*, *false positive**, *false negative**, *specificit**, *screening*, *accuracy*, *predictive value**, *reference value**, *likelihood ratio*', *sroc*, and *receiver operat* characteristic**. We also identified papers that had cited the key papers. Complete details of the search strategy are provided elsewhere (3). We contacted methodologic experts and groups conducting work in this field. Reference lists of retrieved articles were screened for additional studies.

Inclusion Criteria

All studies with the main objective of addressing bias or variation in the results of diagnostic accuracy studies were eligible for inclusion. Studies of any design, including reviews, and any topic area were eligible. Studies had to investigate the effects of bias or variation on measures of test performance, such as sensitivity, specificity, predictive

values, likelihood ratios, and diagnostic odds ratios, and indicate how a particular feature may distort these measures. Inclusion was assessed by one reviewer and checked by a second reviewer; discrepancies were resolved through discussion.

Data Extraction

One reviewer extracted data and a second reviewer checked data on the following parameters: study design, objective, sources of bias or variation investigated, and the results for each source. Discrepancies were resolved by consensus or consultation with a third reviewer.

Data Synthesis

We divided the different sources of bias and variation into groups (**Table 1**). **Table 1** provides a brief description of each source of bias and variation; more detailed descriptions are available elsewhere (3). Results were stratified according to the source of bias or variation. Studies were grouped according to study design. We classified studies that used actual data from one or more clinical studies to demonstrate the effect of a particular study feature as experimental studies, diagnostic accuracy studies, or systematic reviews. Experimental studies were defined as studies specifically designed to test a hypothesis about the effect of a certain feature, for example, rereading sets of radiographs while controlling (manipulating) the overall prevalence of abnormalities. Studies that used models to simulate how certain types of biases may affect estimates of diagnostic test performance were classified as modeling studies. These studies were considered to provide theoretical evidence of bias or variation.

Role of the Funding Source

The funding source was not involved in the design, conduct, or reporting of the study or in the decision to submit the manuscript for publication.

Data Synthesis

The literature searches identified a total of 8663 references. Of these, 569 studies were considered potentially relevant and were assessed for inclusion; 55, published from 1963 to 2000, met inclusion criteria. Nine studies were systematic reviews, 16 studies used an experimental design, 22 studies were diagnostic accuracy studies, and 8 studies used modeling to investigate the theoretical effects of bias or variation.

Population

Demographic Features

Ten studies assessed the effects of demographic features on test performance (**Table 2**) (4, 5, 7, 9, 11, 14, 15, 20, 22, 24). Eight studies were diagnostic accuracy studies, and 2 were systematic reviews. All but one study (22) found an association between the features investigated and overall accuracy. The study that did not find an association investigated whether estimates of exercise testing performance differed between men and women; after correction for the effects of verification bias, no significant differences were found (22).

In general, the studies found associations between the demographic factors investigated and sensitivity; the reported effect on specificity was less strong. Four studies found that various factors, including sex, were associated with sensitivity but showed no association with specificity (4, 5, 11, 20). The index tests investigated in these studies were exercise testing (5, 11, 20) to diagnose heart disease and body mass index to test for obesity (4). Two additional studies of exercise testing also reported an association with sensitivity, but the effects on specificity differed. One found that factors that lead to increased sensitivity also lead to a decrease in specificity (14); the second reported higher sensitivity and specificity in men than in women (16). A study of the diagnostic accuracy of an alcohol screening questionnaire found that overall accuracy was increased in certain ethnic groups (24). Sex was the most commonly investigated variable. Three studies found no association between test performance and sex, 9 found significant effects on sensitivity, and 4 found significant effects on specificity. Other variables shown to have significant effects on test performance were age, race, and smoking status.

Disease Severity

Six studies looked at the effects of disease severity on test performance (**Table 2**) (5, 11, 14, 19, 23, 25). Three studies were diagnostic accuracy studies, 2 were reviews, and one used modeling to investigate the effects of differences in disease severity. The modeling study also included an example from a diagnostic accuracy study of tests for the diagnosis of ovarian cancer (25). Three studies investigated tests for heart disease (5, 11, 14), one examined ventilation–perfusion lung scans for diagnosing pulmonary embolism (23), and one investigated 2 different laboratory tests (one for cancer and the other for bacterial infections) (19). All 6 studies found increased sensitivity with more severe disease; 5 found no effect on specificity (5, 11, 14, 19, 23), and one did not comment on the effects on specificity (25).

Disease Prevalence

Six studies looked at the effects of increased disease prevalence on test performance (**Table 2**) (8, 10, 13, 17, 21, 26). One study used an experimental design (8); the other studies were all diagnostic accuracy studies. The tests investigated in these studies covered a wide range of topics: dipstick for diagnosing urinary tract infection (10), magnetic resonance imaging and evoked potentials for diagnosing multiple sclerosis (17), exercise testing for diagnosing coronary artery disease (21), lung scans for diagnosing pulmonary embolism (8), clinical indications for diagnosing pneumonia (13), and ultrasonography for diagnosing epididymitis (26). Only 5 of the studies reported on the effects of disease prevalence on sensitivity; all found an in-

Table 2. **Population***

Study, Year (Reference)	Design	Index Test	Study Sample
Curtin et al., 1997 (4)	Diagnostic accuracy	Body mass index	226 white persons
Detrano et al., 1988 (5) Detrano et al., 1988 (6)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Egglin and Feinstein, 1996 (8)	Experimental	Pulmonary arteriography	24 arteriograms
Hlatky et al., 1984 (9)	Diagnostic accuracy	Exercise electrocardiography	2269 patients
Lachs et al., 1992 (10)	Diagnostic accuracy	Dipsticks	366 consecutive patients
Levy et al., 1990 (11)	Diagnostic accuracy	Electrocardiography	4684 patients with suspected left ventricular hypotrophy
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Melbye and Straume, 1993 (13)	Diagnostic accuracy	Clinical cues	581 patients with suspected pneumonia
Moons et al., 1997 (14)	Diagnostic accuracy	Exercise test	295 consecutive patients with heart pain
Morise and Diamond, 1994 and 1995 (15, 16)	Diagnostic accuracy	Exercise electrocardiography	4467 patients with suspected coronary disease
O'Connor et al., 1996 (17)	Diagnostic accuracy	Magnetic resonance imaging and evoked potentials	303 patients with suspected multiple sclerosis
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients evaluated for coronary arterial disease
Ransohoff and Feinstein, 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Roger et al., 1997 (20)	Diagnostic accuracy	Exercise echocardiography	3679 consecutive patients
Rozanski et al., 1983 (21)	Diagnostic accuracy	Exercise radionuclide ventriculography	77 angiographically normal patients
Santana-Boado et al., 1998 (22)	Diagnostic accuracy	Single-photon emission computed tomography	702 consecutive patients evaluated for coronary disease
Stein et al., 1993 (23)	Diagnostic accuracy	Ventilation/perfusion scan	1050 patients
Steinbauer et al., 1998 (24)	Diagnostic accuracy	Screening tests for alcohol abuse	1333 adult family practice patients
Taube and Tholander, 1990 (25)	Modeling and diagnostic accuracy	Tests for epithelial ovarian cancer	168 patients with ovarian carcinoma
van der Schouw et al., 1995 (26)	Diagnostic accuracy	Ultrasonography	483 consecutive patients; 372 included
Van Rijkom et al., 1995 (27)	Review	Tests for approximal caries	39 sets of sensitivity and specificity data

* NA = not applicable; ↑ = increased; ↓ = decreased.

crease in sensitivity with increased disease prevalence (8, 10, 13, 17, 26). These studies also investigated the effects of increased disease prevalence on specificity and found mixed results; 2 found that specificity decreased (10, 13), 2 found no effect (8, 17), and one reported increased specificity (26). The remaining study looked only at the effects of disease prevalence on specificity, which was found to decrease (21).

Distorted Selection of Participants

Four studies examined the effects of distorted selection of participants on test performance (**Table 2**) (5, 12, 18, 27). A diagnostic accuracy study of exercise testing for heart disease found that overall accuracy was overestimated if reasons for exclusion commonly used by researchers were applied (18). The other 3 studies were reviews. The first, a review of the clinical and radiologic diagnosis of caries, found that in vivo studies gave higher estimates of test performance than in vitro studies (27). A review of exercise testing for heart disease found that avoiding a limited challenge group (that is, including patients with other confounding diseases or patients taking medications thought to produce false-positive results) did not have significant

Table 2—**Continued**

Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Demographic features	Increased weight; sex (female)	↑	None	NA
Demographic features	Sex, age, and medication use	Associated	None	NA
Distorted selection of participants	Avoidance of limited challenge group	None	None	
Disease severity	Inclusion of patients with previous myocardial infarction	↑	None	NA
Demographic features	Various patient-related characteristics (all are not associated)	Associated	Associated	NA
Disease prevalence	Context of interpretation: effect of increased disease prevalence	↑	None	NA
Demographic features	Exercise heart rate, number of diseased arteries, type of angina, age, and sex	Associated	Associated	NA
Disease prevalence	High pretest probability of disease	↑	↓	NA
Demographic features	Sex (male), increased age, decreased body mass index, not smoking	↑	None	NA
Disease severity	Increased severity of left ventricular hypertrophy	↑	None	NA
Distorted selection of participants	Diagnostic case–control studies	NA	NA	↑
	Nonconsecutive patient enrollment	NA	NA	None
	Retrospective study design	NA	NA	None
	Failure to describe patient spectrum	NA	NA	↑
Disease prevalence	Increased prevalence	↑	↓	NA
Demographic features	Sex, workload, diabetes, smoking, cholesterol level (all are not associated)	↑	↓	NA
Disease severity	Number of diseased vessels	↑	None	NA
Demographic features	Men	↑	↑	NA
Disease prevalence	Increased prevalence	↑	None	NA
Distorted selection of participants	Exclusion of patients with other clinical conditions	NA	NA	↑
Disease severity	Extensive disease	↑	None	NA
Demographic features	Sex (male)	↑	None	NA
Disease prevalence	Increased prevalence	Not reported	↓	NA
Demographic features	Sex	None	None	NA
Disease severity	Previous pulmonary disease	↑	None	NA
Demographic features	Race and sex	NA	NA	Associated
Disease severity	Clear cases of malignant disease	↑	Not reported	NA
Disease prevalence	Increased prevalence (inclusion criteria widened)	↑	↑	NA
Distorted selection of participants	In vivo studies compared with in vitro studies	NA	NA	↑

effects on overall accuracy (5). The final study, which reviewed many different tests, found that case–control studies overestimate overall accuracy; it also found that nonconsecutive patient enrollment and a retrospective study design did not affect the diagnostic odds ratio (12). This review also looked at the effects of failure to provide an appropriate description of the patient sample and found that this was associated with increased overall accuracy.

Test Protocol: Materials and Methods

Test Execution

We found only 2 studies, both reviews, that specifically looked at the effects of differences in test execution (**Table 3**) (6, 12). The first, a review of several different tests, found that failure to describe the index test and reference standard execution leads to an overestimation of overall accuracy (12). The other found no effect of differences in protocol on overall accuracy in exercise testing (6).

Test Technology

Two studies looked at the effects of a change in the technology of the index test on test performance (**Table 3**) (6, 28). A systematic review of exercise scintigraphy studies found that automation of the test procedure improved sensitivity but decreased specificity (6). The other study, a

Table 3. **Test Protocol: Materials and Methods***

Study Details	Design	Index Test	Study Sample	Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Detrano et al., 1988 (6)	Review	Exercise electrocardiography	60 primary studies	Test execution Test technology Disease progression bias	Exercise protocol Automation of test Maximum interval between scintigraphy and angiography	None ↑ None	None ↓ None	NA NA NA
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina pectoris	Test technology	Computerized readings	None	None	NA
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests	Test execution	Failure to describe index test execution; failure to describe reference standard execution	NA NA	NA NA	↑ ↓

* NA = not applicable; ↑ = increased; ↓ = decreased.

diagnostic accuracy study of the electrocardiographic exercise test, found no effect on test performance (28).

Treatment Paradox and Disease Progression Bias

No studies that provided evidence of the effect of treatment paradox were identified. Only one study that looked at the effects of disease progression bias on test performance was found. This study, a review of exercise scintigraphy for the diagnosis of heart disease, found no evidence of bias (6).

Reference Standard and Verification Procedure

Inappropriate Reference Standard

Eight studies looked at reference standard error bias (**Table 4**) (6, 7, 27, 29, 31, 34, 41, 43). Four were systematic reviews, and the other 4 used modeling to investigate the theoretical effects of an imperfect reference standard. The reviews looked at reference standard error bias from slightly different perspectives, but all found evidence of bias. A review of patients who received a diagnosis of caries found that weaker validation methods may overestimate overall accuracy (27). A review of a hormone test for the diagnosis of depression found that different reference standards can provide very different estimates of sensitivity (29). A review of exercise scintigraphy for the diagnosis of heart disease found that studies that used a specific reference standard (tomographic imaging) overestimated sensitivity and specificity compared with other studies (6). The last review, which dealt with exercise electrocardiography for heart disease, found that comparison with a more accurate test leads to increased sensitivity but did not report on the effect on specificity (7).

The studies that used modeling to investigate the effects of an imperfect reference standard also found evidence of bias. One study suggested that with imperfect reference standards, specificity is most accurately estimated at low disease prevalence and sensitivity at high disease prevalence; it also suggested that considerable errors in estimates exist, even when the reference standard has close to perfect performance (31). Two studies found that inaccurate reference standards lead to underestimation of index test accuracy when the index test errors are statistically independent of the reference standard and overestimation when the index test errors are statistically dependent on the reference standard (41, 43). The final study found that overall accuracy is underestimated when the test being evaluated is more accurate than the reference standard (34, 43).

Differential Verification Bias

Only 2 studies looked at differential verification bias (**Table 4**) (12, 30). One was a review of several different tests (12), and the other was a diagnostic accuracy study of the clinical diagnosis of Alzheimer disease (30). Both found that differential verification bias leads to higher (inflated) measures of overall accuracy.

Partial Verification Bias

Twenty studies investigated the effects of partial verification bias (**Table 4**) (5, 7, 12, 16, 18–22, 28, 30, 32, 35–40, 42, 44). Two studies used models to investigate the theoretical effects of verification bias and found that partial verification bias increased sensitivity and decreased specificity (35, 36). A third study also used modeling to investigate the effects of verification bias; in addition, it provided an example from a diagnostic accuracy study. This study reported an association between overall accuracy and the presence of partial verification bias (44).

All of the remaining studies used actual data to investigate the effects of partial verification bias and were either diagnostic accuracy studies or reviews. Most of these studies examined some form of exercise testing for the diagnosis of heart disease (5, 6, 16, 18, 20, 21, 28, 32, 38). Other tests that were investigated included noninvasive tests for arterial disease (37), clinical diagnosis for Alzheimer disease (30), clinical findings for diagnosing hemorrhage in patients who had strokes (40), nuchal translucency for diagnosing Down syndrome (39), the carcinoembryonic antigen and nitro-blue tests (19), and serum ferritin levels for diagnosing hereditary hemochromatosis (42). Seven studies

found that sensitivity was increased and specificity decreased in the presence of partial verification bias (16, 18, 20, 28, 32, 38, 40); one study found that both sensitivity and specificity were increased (39), and 2 studies found that sensitivity was increased but did not report the effects on specificity (19, 42). One study found that specificity was increased in the presence of verification bias (5) and another study reported that verification bias decreased specificity (21). Neither of these studies reported on the effects on sensitivity. Two studies did not report on the effects of partial verification bias on sensitivity and specificity. One of these found that partial verification bias increased overall accuracy (37), and the second reported that there was "scope for verification bias" but provided no additional information (30).

Two more studies found no evidence of bias. One was a systematic review of studies of the diagnostic accuracy of exercise electrocardiography (45), and the other was a review of systematic reviews of several different tests (12). The latter study used the relative diagnostic odds ratio as the summary statistic. If partial verification bias tends to increase sensitivity and decrease specificity, as is suggested by some of the studies, then no effect on the diagnostic odds ratio would be expected. This may explain why this review did not find any evidence of partial verification bias.

Interpretation (Reading Process)

Review Bias

Four studies investigated review bias (6, 12, 19, 45), 3 (6, 19, 45) examined diagnostic and test review bias, and one looked only at diagnostic review bias (**Table 5**) (12). A review of exercise testing found no effect of either diagnostic or test review bias on sensitivity and specificity (7). A separate review of exercise testing reported that both diagnostic and test review bias led to an increase in sensitivity but had no effect on specificity (5). A study of carcinoembryonic antigen and nitro-blue tests found that failure to avoid review bias may overestimate sensitivity and specificity (19). A review of several different tests looked only at diagnostic review bias and found that it increased overall accuracy (12).

Clinical Review Bias

Nine studies looked at the effects of clinical review bias (**Table 5**) (28, 46, 52, 53, 55–57, 59, 61). Most of these studies examined radiography (46, 52, 56, 57, 61), mammography (55), and myelography and spinal computer tomography (53). Eight studies used an experimental design, and one was a diagnostic accuracy study (28). One found no difference in overall accuracy between tests interpreted with and without clinical history (56). The other studies all found evidence of bias; however, the direction of bias differed among studies. In general, studies found that providing clinical information improved overall accuracy. Six studies reported that sensitivity was increased when clinical information was available (28, 46, 52, 53, 57, 61). The effects of providing clinical information on specificity varied among these studies: Two reported that specificity decreased (52, 53), 2 found no effect on specificity (46, 61), and the other 2 did not report on the effects on specificity (28, 57). The remaining 2 studies did not report on the effects of providing clinical history on sensitivity and specificity, but both found that overall accuracy was improved when clinical information was provided (55, 59).

Incorporation Bias

No studies that looked at the effects of incorporation bias were identified.

Observer Variability

Eight studies looked at observer variation; no studies addressed instrument variation (**Table 5**) (47–51, 54, 58, 60). All studies used an experimental design. Most studies were evaluations of imaging techniques: radiologic detection of fractures (47), mammography (48, 54), and myocardial imaging (51). Other techniques that were evaluated were fine-needle aspiration biopsy (49), histologic examination (50), cytologic examination (60), and bronchial brush specimens (58). All 8 studies found evidence of interobserver variability, and 2 found evidence of intra-observer variability (48, 50); one of these studies reported that interobserver variability was greater than intraobserver variability (48). Two studies found that more experienced reviewers, or experts, provided greater sensitivity (49, 60), whereas another found that experience was not related to interobserver variability (58).

Analysis

Handling of Indeterminate Results

Two studies looked at the effects of uninterpretable test results (**Table 6**) (7, 18). One of these studies stated that a large proportion of results would be excluded if unsatisfactory test results were excluded but provided no evidence on how this may lead to biased estimates of test performance (18). The other study found that the treatment of equivocal or nondiagnostic test results was not associated with overall accuracy (7).

Arbitrary Choice of Threshold Value

No studies that provided evidence of the effect of the choice of threshold value were identified.

Discussion

The searches identified a relatively small number of studies that looked specifically at the effects of bias and variation on estimates of diagnostic test performance. These studies were concentrated in 7 areas of bias and variation: demographic features (10 studies), disease prevalence (6 studies), disease severity (6 studies), inappropriate reference standard (8 studies), partial verification bias (20 studies), clinical review bias (9 studies), and observer variation (8 studies). Other sources of bias commonly believed

Table 4. **Reference Standard and Verification Procedure***

Study Details	Design	Index Test	Study Sample
Arana et al., 1990 (29)	Review	Thyrotropin-releasing hormone stimulation	10 studies
Bowler et al., 1998 (30)	Diagnostic accuracy	Necropsy	307 patients
Boyko et al., 1988 (31)	Modeling	NA	Formulas used to model theoretical effects
Cecil et al., 1996 (32)	Diagnostic accuracy	Stress single-photon emission computed tomography thallium testing	4354 records selected from computerized database
De Neef, 1987 (34)	Modeling	New rapid antigen detection tests	Models used to vary reference standard accuracy
Detrano et al., 1988 (5, 6)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Diamond, 1991 (35)	Modeling	NA	Series of computer simulations using the Begg–Greenes method†
Diamond, 1992 (36)	Modeling	NA	Series of computer simulations using Bayes theorem
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Lijmer et al., 1996 (37)	Diagnostic accuracy	Noninvasive tests	464 consecutive patients with suspected disease
Miller et al., 1998 (38)	Diagnostic accuracy	Stress imaging	15 945 low-risk patients
Mol et al., 1999 (39)	Review	Nuchal translucency measurement	25 studies
Morise and Diamond, 1994 and 1995 (15, 16)	Diagnostic accuracy	Exercise electrocardiography	4467 patients with suspected coronary disease
Panzer et al., 1987 (40)	Diagnostic accuracy	Clinical findings	374 patients with stroke and focal deficits
Phelps and Hutson, 1995 (41)	Modeling	NA	Monte Carlo studies
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients
Ransohoff and Muir, 1982 (42)	Review	Serum ferritin levels	2 studies
Ransohoff et al., 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Roger et al., 1997 (20)	Diagnostic accuracy	Exercise echocardiography	3679 consecutive patients
Rozanski et al., 1983 (21)	Diagnostic accuracy	Exercise ventriculography	77 angiographically normal patients
Santana-Boado et al., 1998 (22)	Diagnostic accuracy	Single-photon emission computed tomography	702 consecutive low-risk patients
Thibodeau, 1981 (43)	Modeling	NA	Various statistical models
van Rijkom and Verdonschot, 1995 (27)	Review	Tests for approximal caries	39 sets of sensitivity and specificity data
Zhou, 1994 (44)	Modeling and diagnostic accuracy	NA	429 patients

* DSM-III = *Diagnostic and Statistical Manual of the American Psychological Association*, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; ↑ = increased; ↓ = decreased.
† From Begg C and Greenes R (33).

to affect studies of diagnostic test performance, such as incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts, were not considered in any studies.

Population

The evidence shows that differences in populations affect estimates of diagnostic performance. However, the extent and direction of the effect of variations in a population can vary, even among studies of the same index test. Demographic features have shown strong associations with test performance and generally showed a greater effect on estimates of sensitivity than on specificity. Studies that observed effects on specificity generally found that factors that increased sensitivity also decreased estimates of specificity. There was also evidence that both disease severity and prevalence may affect estimates of test performance. Sensitivity tended to be increased in populations with more

Table 4—**Continued**

Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Inappropriate reference standard	DSM-III instead of RDC as the reference test	↓	Not reported	NA
Differential and partial verification bias	Autopsy to confirm the clinical diagnosis	NA	NA	"Scope for bias"
Inappropriate reference standard	Effects of reference standard errors	NA	NA	Associated
Partial verification bias	Effect of partial verification bias using the Begg method†	↑	↓	NA
Inappropriate reference standard	Increased sensitivity of the reference standard	↑	Large errors	NA
Inappropriate reference standard	Tomographic imaging instead of angiography as reference test	↑	↑	NA
Partial verification bias	Presence of partial verification bias	None	↑	NA
Inappropriate reference standard	Exercise test thought to be superior in accuracy as reference standard	Associated	Not reported	NA
Partial verification bias	Presence of partial verification bias	NA	NA	None
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Differential verification bias	Studies that used different reference standards	NA	NA	↑
Partial verification bias	Presence of partial verification bias	NA	NA	None
Partial verification bias	Presence of partial verification bias	NA	NA	↑
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	↑	↑	NA
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Inappropriate reference standard	Use of inaccurate "fuzzy" reference standard	NA	NA	Associated
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	↑	Not reported	NA
Partial verification bias	Presence of partial verification bias	↑	Not reported	NA
Partial verification bias	Presence of partial verification bias	↑	↓	NA
Partial verification bias	Presence of partial verification bias	Not reported	↓	NA
Partial verification bias	Presence of partial verification bias	None	None	NA
Inappropriate reference standard	Use of inaccurate reference standard	NA	NA	Associated
Inappropriate reference standard	Use of weak validation methods	NA	NA	↑
Partial verification bias	Presence of partial verification bias	NA	NA	Associated

severe disease or increased disease prevalence. Disease severity had little effect on estimates of specificity, and the effect of disease prevalence on specificity varied. The way in which participants are selected for inclusion in studies of diagnostic accuracy has also been shown to affect test performance. However, the studies that investigated this variable looked at very different aspects of patient selection; thus, it is difficult to draw overall conclusions.

Test Protocol

Very few studies investigated the effects of biases and sources of variation associated with test protocol, and those that did reported mixed results. Because of the lack of evidence on the effects of test protocol, it is difficult to draw conclusions regarding the effect of this variable on estimates of test performance. The magnitude of the effect of these biases and sources of variation is probably linked to the test and condition being investigated. For example, the effect of differences in test execution is probably much greater for a test that requires some degree of expertise to perform than for a test that is very straightforward to perform. Similarly, treatment paradox and disease progression bias are more likely to have significant effects on studies of

Table 5. **Interpretation (Reading Process)***

Study Details	Design	Index Test	Study Sample
Arana et al., 1990 (29)	Review	Thyrotropin-releasing hormone stimulation	10 studies
Berbaum et al., 1988 (46)	Experimental	Radiography	40 radiographs examined with and without clinical information
Berbaum et al., 1989 (47)	Experimental	Radiography	40 radiographs examined by a group of radiologists and a group of orthopedic surgeons
Ciccone et al., 1992 (48)	Experimental	Mammography	45 mammograms; 7 radiologists
Cohen et al., 1987 (49)	Experimental	Fine-needle aspiration biopsy	50 specimens examined by 5 observers
Corley et al., 1997 (50)	Experimental	Histologic diagnosis of pneumonia	39 lung biopsy samples, 4 pathologists
Cuaron et al., 1980 (51)	Experimental	Tc 99m phosphate myocardial imaging	250 myocardial slides evaluated by 6 observers
Detrano et al., 1988 (5, 6)	Review	Exercise thallium scintigraphy	56 primary studies
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies
Doubilet et al., 1981 (52)	Experimental	Radiography	8 test radiographs; 4 with suggestive and 4 nonsuggestive history
Eldevik et al., 1982 (53)	Experimental	Myelography and computed tomography	107 patients assessed with and without clinical history
Elmore et al., 1994 (54)	Experimental	Mammography	150 mammograms, 10 radiologists
Elmore et al., 1997 (55)	Experimental	Mammography	100 radiographs assessed with and without clinical history
Froelicher et al., 1998 (28)	Diagnostic accuracy	Electrocardiography and angiographic calipers	814 consecutive patients with angina
Good et al., 1990 (56)	Experimental	Chest radiography	247 radiographs assessed with and without clinical history
Lijmer et al., 1999 (12)	Review	Various tests	184 primary studies of 218 tests
Potchen et al., 1979 (57)	Experimental	Chest radiography	3 groups of radiologists; different combinations of data
Raab et al., 1995 (58)	Experimental	Bronchial brush specimens	100 bronchial brush specimens examined by different observers
Raab et al., 2000 (59)	Experimental	Bronchial brush specimens	97 specimens, assessed with and without clinical information
Ransohoff et al., 1978 (19)	Review	Carcinoembryonic antigen and nitroblue tetrazolium tests	17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium
Ronco et al., 1996 (60)	Experimental	Colpohistologic and cytologic screening	61 samples examined by cytologists and experts
Schreiber, 1963 (61)	Experimental	Chest radiography	100 chest radiographs assessed with and without clinical information

* DSM-III = *Diagnostic and Statistical Manual of the American Psychological Association*, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; ↑ = increased; ↓ = decreased.

tests for acute diseases that may be easily treated (for example, infections) and that may change more rapidly than chronic conditions that do not respond well to treatment and that may remain in the same stage for longer periods.

Reference Standard

The evidence was strong for the effect of biases associated with verification procedure on test performance. All studies that looked at the effects of using an inappropriate reference standard found that test performance was affected; however, the direction of the effect differed among studies. Theoretically, if the reference standard is not 100% accurate, the index test may correctly classify results that have been incorrectly classified by the reference standard. This would be expected to lead to an underestimation of test performance. It is also possible that an imperfect reference standard may classify results of the index test as being correct when they are actually incorrect. This would be expected to lead to overestimation of test performance. Thus, an inaccurate reference standard could affect test performance in either way.

Many studies looked at the effects of verification bias, especially partial verification bias. Most reported that verification influenced estimates of test performance. In theory, if all of the patients with negative test results are not verified by the reference standard and are subsequently omitted from the 2×2 table, estimates of sensitivity would be inflated because patients with false-negative test results will go undetected. This is supported by the evidence; all studies that observed a significant effect on sensitivity found that sensitivity was increased in the presence of verification bias. However, as with many other biases, the effects on specificity were less clear.

Table 5—**Continued**

Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Inappropriate reference standard	DSM-III instead of RDC as the reference standard	↓	Not reported	NA
Clinical review bias	Availability of clinical information	↑	None	↑
Observer variation	Difference between radiologists and orthopedic surgeons	NA	NA	Associated
Observer variation	Difference between radiologists and orthopedic surgeons	NA	NA	Associated
Observer variation	Inter- and intraobserver variation	NA	NA	Associated
Observer variation	Effect of training and experience	↑	↑	NA
Observer variation	Inter- and intraobserver variation	NA	NA	None
Observer variation	Interobserver variation	NA	NA	Associated
Review bias	Lack of blinding, that is, presence of review bias	↑	Not reported	↑
Review bias	Lack of blinding, that is, presence of review bias	NA	NA	None
Clinical review bias	Suggestive clinical history	↑	↓	NA
Clinical review bias	Availability of clinical information	↑	↓	NA
Observer variation	Interobserver variation	NA	NA	Associated
Clinical review bias	Availability of clinical information	NA	NA	↑
Clinical review bias	Availability of clinical information	↑	Not reported	NA
Clinical review bias	Availability of clinical information	NA	NA	None
Review bias	Lack of blinding, that is, presence of review bias	NA	NA	↑
Clinical review bias	Availability of clinical information	↑	Not reported	NA
Observer variation	Interobserver variation	NA	NA	Associated
Clinical review bias	Availability of clinical information	NA	NA	↑
Review bias	Lack of blinding, that is, presence of review bias	↑	↑	NA
Observer variation	Effect of training and experience (being an "expert")	↑	Not reported	NA
Clinical review bias	Availability of clinical information	↑	None	NA

Interpretation

Reading processes that involve interpretation of results affect estimates of test performance. Both diagnostic and test review biases were found to increase sensitivity; however, no effect on specificity was noted. An effect on sensitivity would be expected because knowledge of the index test result when interpreting the reference standard (or vice versa) probably increases the agreement between tests. This in turn leads to a greater number of true-positives and true-negative results and would be expected to increase estimates of both sensitivity and specificity. It is unclear why studies did not find significant effects on specificity. Perhaps the effects on specificity are smaller and any effect may therefore not reach statistical significance.

The availability of clinical information to the person interpreting the results of the index test was found to increase sensitivity. Although the evidence for an effect on specificity was minimal, specificity decreased in 2 studies. The provision of clinical information probably has different effects depending on the test being evaluated. Whether clinical information should be available in a particular diagnostic study should be carefully considered in each case. It seems that the best approach to interpreting the results of a diagnostic accuracy study would be to determine whether the clinical information available to those interpreting the results of the index test is the same as the clinical information that would be available when the test is interpreted in practice.

All studies that looked at the effects of observer variation found significant differences among observers in their estimates of test performance. Therefore, the effects of observer variation will inevitably be greater for tests that involve a strong degree of subjective interpretation compared with a fully automated test.

Table 6. **Analysis***

Study Details	Design	Index Test	Study Sample	Source of Bias or Variation	Factors Investigated	Effect on Sensitivity	Effect on Specificity	Effect on Overall Accuracy
Detrano et al., 1989 (7)	Review	Exercise electrocardiography	60 primary studies	Handling of indeterminate results	Treatment of equivocal or nondiagnostic tests	NA	NA	None
Philbrick et al., 1982 (18)	Diagnostic accuracy	Graded exercise test	208 consecutive patients	Handling of indeterminate results	Exclusion of unsatisfactory exercise test results	NA	NA	Unclear

* NA = not applicable.

Analysis

Very few studies investigated the effects of biases associated with analysis on test performance. The effect of the exclusion of indeterminate results and the nonarbitrary choice of threshold value remains unclear from the evidence reviewed.

Limitations

The main limitation of our review is the difficulty in identifying articles that examined specific features of the design and conduct of diagnostic studies. Indexing on MEDLINE and other electronic databases focuses on diseases, therapies, and test technologies and not on elements of design. There is no specific way of indexing studies that relate to the diagnostic accuracy of a test (1). In addition, many different names have been used to label the same phenomenon in studies of diagnostic accuracy tests. To try to overcome these difficulties, very broad searches were performed. However, we may have still missed several relevant papers. The information provided in our paper should provide useful examples but may not be comprehensive.

Ideally, we would have liked to provide a quantitative synthesis to assess the magnitude of each of the biases and sources of variation as well as their direction. However, because the studies included were very heterogeneous, a quantitative synthesis was not possible. The studies also measured the effect of the biases and sources of variation in different ways. In particular, diagnostic accuracy and experimental studies looked at the effect of biases and sources of variation within studies, whereas reviews looked at reasons for differences in estimates among studies. It is also likely that different biases and sources of variation will be important in different topic areas. For example, observer variation is likely to be a problem only for studies that involve some degree of subjective interpretation. Also, observer variation is likely to have a greater effect with more subjective interpretations.

Another problem is that sources of bias and variation may act differently depending on the study. For example, for partial verification bias, the effects may differ when the reference standard is not used in selected groups. The group that does not receive verification may, for example, be a random sample of patients, a selected subgroup of patients with negative test results, or all patients with positive test results. All of these situations are called partial verification, but the effects of each situation probably differ. Within a single study, there is only one true effect of a feature, but this true effect may differ depending on the study. Chance and the effect of other factors may obscure the true effect. These factors combine to create difficulty in determining the overall effect of a source of bias or variation.

We included studies that provided both real-life examples of the effects of different biases and sources of variation as well as studies that used modeling to investigate the effects of different biases or sources of variation. When the results of the modeling studies are interpreted, it is important to consider that these studies can provide an indication only of the theoretical effect of a source of bias or variation. The results from these studies need to be supported by additional empirical evidence from real-life examples before more firm conclusions can be drawn (12).

Conclusions

This paper provides information on the available evidence for the effects of each source of bias and variation in diagnostic accuracy studies. The sources of bias and variation for which there is the most evidence are demographic features, disease prevalence or severity, partial verification bias, clinical review bias, and observer or instrument variation. Some evidence was also available for the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias. The potential effects of these biases and sources of variation should be considered when interpreting or designing diagnostic accuracy studies. Additional research should be done to investigate potential sources of bias and variation.

From the University of York, York, United Kingdom, and the University of Amsterdam, Amsterdam, the Netherlands.

Disclaimer: The views expressed in this paper are those of the authors and not necessarily those of the Standing Group, the Commissioning Group, or the Department of Health.

Acknowledgments: The authors thank Kath Wright (Centre for Reviews and Dissemination) for carrying out literature searches. They also

thank the advisory panel to the review for their help during various stages, including commenting on the protocol and draft report.

Grant Support: Commissioned and funded by the National Health Service R&D Health Technology Assessment Programme (project number 98/27/99).

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Penny Whiting, MSc, Centre for Reviews and Dissemination, University of York, York YO10 5DD, United Kingdom; e-mail, pfw2@york.ac.uk.

Current author addresses are available at www.annals.org.

References

1. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.** The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003;49:7-18. [PMID: 12507954]

2. **Feinstein AR.** Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia: WB Saunders Co; 1985.

3. **Whiting PJ, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J.** The development and validation of methods for assessing the quality and reporting of diagnostic studies. Health Technol Assess [In press].

4. **Curtin F, Morabia A, Pichard C, Slosman DO.** Body mass index compared to dual-energy x-ray absorptiometry: evidence for a spectrum bias. J Clin Epidemiol. 1997;50:837-43. [PMID: 9253396]

5. **Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher VF.** Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med. 1988;84:699-710. [PMID: 3041808]

6. **Detrano R, Lyons KP, Marcondes G, Abbassi N, Froelicher VF, Janosi A.** Methodologic problems in exercise testing research. Are we solving them? Arch Intern Med. 1988;148:1289-95. [PMID: 3288157]

7. **Detrano R, Gianrossi R, Mulvihill D, Lehmann K, Dubach P, Colombo A, et al.** Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: a meta analysis. J Am Coll Cardiol. 1989;14:1501-8. [PMID: 2809010]

8. **Egglin TK, Feinstein AR.** Context bias. A problem in diagnostic radiology. JAMA. 1996;276:1752-5. [PMID: 8940325]

9. **Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA.** Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med. 1984;77:64-71. [PMID: 6741986]

10. **Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS.** Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med. 1992;117:135-40. [PMID: 1605428]

11. **Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP.** Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation. 1990;81:815-20. [PMID: 2137733]

12. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]

13. **Melbye H, Straume B.** The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. Scand J Prim Health Care. 1993; 11:241-6. [PMID: 8146507]

14. **Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE.** Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12-7. [PMID: 9116087]

15. **Morise AP, Diamond GA.** Does sex discrimination explain the differences in test accuracy among men and women referred for exercise electrocardiography? Circulation. 1994;90(Pt 2):1-273.

16. **Morise AP, Diamond GA.** Comparison of the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women. Am Heart J. 1995;130:741-7. [PMID: 7572581]

17. **O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W.** The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology. 1996;47: 140-4. [PMID: 8710067]

18. **Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP.** The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA. 1982;248:2467-70. [PMID: 7131702]

19. **Ransohoff DF, Feinstein AR.** Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30. [PMID: 692598]

20. **Roger VL, Pellikka PA, Bell MR, Chow CW, Bailey KR, Seward JB.** Sex and test verification bias. Impact on the diagnostic value of exercise echocardiography. Circulation. 1997;95:405-10. [PMID: 9008457]

21. **Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ.** The declining specificity of exercise radionuclide ventriculography. N Engl J Med. 1983;309:518-22. [PMID: 6877322]

22. **Santana-Boado C, Candell-Riera J, Castell-Conesa J, Aguada-Bruix S, Garcia-Burillo A, Canela T, et al.** Diagnostic accuracy of technetium-99m-MIBI myocardial SPECT in women and men. J Nucl Med. 1998;39:751-5. [PMID: 9591568]

23. **Stein PD, Gottschalk A, Henry JW, Shivkumar K.** Stratification of patients according to prior cardiopulmonary disease and probability assessment based on the number of mismatched segmental equivalent perfusion defects. Approaches to strengthen the diagnostic value of ventilation/perfusion lung scans in acute pulmonary embolism. Chest. 1993;104:1461-7. [PMID: 8222807]

24. **Steinbauer JR, Cantor SB, Holzer CE 3rd, Volk RJ.** Ethnic and sex bias in primary care screening tests for alcohol use disorders. Ann Intern Med. 1998;129: 353-62. [PMID: 9735062]

25. **Taube A, Tholander B.** Over- and underestimation of the sensitivity of a diagnostic malignancy test due to various selections of the study population. Acta Oncol. 1990;29:971-6. [PMID: 2278729]

26. **van der Schouw YT, Van Dijk R, Verbeek AL.** Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. J Clin Epidemiol. 1995;48:417-22. [PMID: 7897462]

27. **van Rijkom HM, Verdonschot EH.** Factors involved in validity measurements of diagnostic tests for approximal caries—a meta-analysis. Caries Res. 1995;29:364-70. [PMID: 8521438]

28. **Froelicher VF, Lehmann KG, Thomas R, Goldman S, Morrison D, Edson R, et al.** The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction. Veterans Affairs Cooperative Study in Health Services #016 (QUEXTA) Study Group. Quantitative Exercise Testing and Angiography. Ann Intern Med. 1998;128:965-74. [PMID: 9625682]

29. **Arana GW, Zarzar MN, Baker E.** The effect of diagnostic methodology on the sensitivity of the TRH stimulation test for depression: a literature review. Biol Psychiatry. 1990;28:733-7. [PMID: 2122917]

30. **Bowler JV, Munoz DG, Merskey H, Hachinski V.** Fallacies in the pathological confirmation of the diagnosis of Alzheimer's disease. J Neurol Neurosurg Psychiatry. 1998;64:18-24. [PMID: 9436722]

31. **Boyko EJ, Alderman BW, Baron AE.** Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. J Gen Intern Med. 1988;3:476-81. [PMID: 3049969]

32. **Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, et al.** The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol. 1996;49:735-42. [PMID: 8691222]

33. **Begg CB, Greenes RA.** Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39:207-15. [PMID: 6871349]

34. **De Neef P.** Evaluating rapid tests for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison. Med Decis Making. 1987;7:92-6. [PMID: 3553828]

35. **Diamond GA.** Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? Med Decis Making. 1991;11:48-56. [PMID: 2034075]

36. **Diamond GA.** Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. Med Decis Making. 1992;12:22-31. [PMID:

1538629]

37. **Lijmer JG, Hunink MG, van den Dungen JJ, Loonstra J, Smit AJ.** ROC analysis of noninvasive tests for peripheral arterial disease. Ultrasound Med Biol. 1996;22:391-8. [PMID: 8795165]

38. **Miller TD, Hodge DO, Christian TF, Milavetz JJ, Bailey KR, Gibbons RJ.** The impact of adjusting for post-test referral bias on apparent sensitivity and specificity of SPECT myocardial perfusion imaging in men and women [Abstract]. J Am Coll Cardiol. 1998;31(2 Suppl A):167A.

39. **Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM.** Effect of study design on the association between nuchal translucency measurement and Down syndrome. Obstet Gynecol. 1999;94:864-9. [PMID: 10546775]

40. **Panzer RJ, Suchman AL, Griner PF.** Workup bias in prediction research. Med Decis Making. 1987;7:115-9. [PMID: 3574021]

41. **Phelps CE, Hutson A.** Estimating diagnostic test accuracy using a "fuzzy gold standard." Med Decis Making. 1995;15:44-57. [PMID: 7898298]

42. **Ransohoff DF, Muir WA.** Diagnostic workup bias in the evaluation of a test. Serum ferritin and hereditary hemochromatosis. Med Decis Making. 1982;2: 139-45. [PMID: 7167042]

43. **Thibodeau L.** Evaluating diagnostic tests. Biometrics. 1981:801-4.

44. **Zhou XH.** Effect of verification bias on positive and negative predictive values. Stat Med. 1994;13:1737-45. [PMID: 7997707]

45. **Detrano R, Gianrossi R, Froelicher V.** The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. Prog Cardiovasc Dis. 1989;32:173-206. [PMID: 2530605]

46. **Berbaum KS, el-Khoury GY, Franken EA Jr, Kathol M, Montgomery WJ, Hesson W.** Impact of clinical history on fracture detection with radiography. Radiology. 1988;168:507-11. [PMID: 3393672]

47. **Berbaum KS, Franken EA Jr, el-Khoury GY.** Impact of clinical history on radiographic detection of fractures: a comparison of radiologists and orthopedists. AJR Am J Roentgenol. 1989;153:1221-4. [PMID: 2816635]

48. **Ciccone G, Vineis P, Frigerio A, Segnan N.** Inter-observer and intra-observer variability of mammogram interpretation: a field study. Eur J Cancer. 1992;28A:1054-8. [PMID: 1627374]

49. **Cohen MB, Rodgers RP, Hales MS, Gonzales JM, Ljung BM, Beckstead JH, et al.** Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. Arch Pathol Lab Med. 1987;111:518-20. [PMID: 3579506]

50. **Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE, et al.** Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. Chest. 1997;112: 458-65. [PMID: 9266884]

51. **Cuaron A, Acero AP, Cardenas M, Huerta D, Rodriguez A, de Garay R.** Interobserver variability in the interpretation of myocardial images with Tc-99m-labeled diphosphonate and pyrophosphate. J Nucl Med. 1980;21:1-9. [PMID: 6243350]

52. **Doubilet P, Herman PG.** Interpretation of radiographs: effect of clinical history. AJR Am J Roentgenol. 1981;137:1055-8. [PMID: 6975000]

53. **Eldevik OP, Dugstad G, Orrison WW, Haughton VM.** The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiology. 1982;145:85-9. [PMID: 7122902]

54. **Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR.** Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994;331:1493-9. [PMID: 7969300]

55. **Elmore JG, Wells CK, Howard DH, Feinstein AR.** The impact of clinical history on mammographic interpretations. JAMA. 1997;277:49-52. [PMID: 8980210]

56. **Good BC, Cooperstein LA, DeMarino GB, Miketic LM, Gennari RC, Rockette HE, et al.** Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? AJR Am J Roentgenol. 1990;154:709-12. [PMID: 2107662]

57. **Potchen E, Gard J, Lazar P, Lahaie P, Andary M.** The effect of clinical history data on chest film interpretation: direction or distraction [Abstract]. Invest Radiol. 1979;14:404.

58. **Raab SS, Thomas PA, Lenel JC, Bottles K, Fitzsimmons KM, Zaleski MS, et al.** Pathology and probability. Likelihood ratios and receiver operating characteristic curves in the interpretation of bronchial brush specimens. Am J Clin Pathol. 1995;103:588-93. [PMID: 7741104]

59. **Raab SS, Oweity T, Hughes JH, Salomao DR, Kelley CM, Flynn CM, et al.** Effect of clinical history on diagnostic accuracy in the cytologic interpretation of bronchial brush specimens. Am J Clin Pathol. 2000;114:78-83. [PMID: 10884802]

60. **Ronco G, Montanari G, Aimone V, Parisio F, Segnan N, Valle A, et al.** Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology. 1996;7:151-8. [PMID: 8782987]

61. **Schreiber M.** The clinical history as a factor in roentgenogram interpretation. JAMA. 1963;185:137-9.

Test Research versus Diagnostic Research

The diagnostic workup starts with a patient presenting with symptoms or signs suggestive of a particular disease. The workup is commonly a consecutive process starting with medical history and physical examination and simple tests followed by more burdensome and costly diagnostic procedures. Generally, after each test all available results are converted (often implicitly) to a probability of disease, which in turn directs decisions for additional testing or initiation of appropriate treatment. Setting a diagnosis is a multitest or multivariable process of estimating and updating the diagnostic probability of disease presence given combinations of test results. Each test may be more or less burdensome to the patient, time-consuming, and/or costly. Different tests often provide to various degrees the same information because they are all associated with the same underlying disorder. Relevant for physicians is to know which tests are redundant and which have true, independent predictive value for the presence or absence of the target disease. Accordingly, studies of diagnostic accuracy should demonstrate which (subsequent) test results truly increase or decrease the probability of disease presence as estimated from the previous results, and to what extent.

Various reviews have demonstrated that the majority of published studies of diagnostic accuracy still have methodologic flaws in design or analysis or provide results with limited practical applicability *(1–3)*. This has been attributed to the absence of a proper methodologic framework for diagnostic test evaluations as, for example, exists for studies of therapies and etiologic factors and has motivated various researchers to establish frameworks for studies of diagnostic accuracy, such as the recent STARD Initiative *(4–12)*. In our view, an issue that has received too little attention in most of these methodologic essays is the difference between test research and diagnostic research.

By "test research" we refer to studies that follow a single-test or univariable approach, i.e., studies focusing on a particular test to quantify its sensitivity, specificity, likelihood ratio (LR), or area under the ROC curve (ROC area). We call this test research because it merely quantifies the "characteristics" of the test rather than the test's contribution to estimate the diagnostic probability of disease presence or absence. By "diagnostic research" we refer to studies that aim to quantify a test's added contribution beyond test results readily available to the physician in determining the presence or absence of a particular disease. Although the multivariable and probabilistic character of medical diagnosis is slowly gaining appreciation in medical research, the majority of studies on diagnostic accuracy may still be regarded as test research *(2, 3, 8)*.

We believe that test research has limited applicability to clinical practice. Below we describe why we believe this is the case, provide a brief description of a better approach, and give two clinical examples illustrating the hazards of test research. Finally, we describe the few instances in which test research may be worthwhile.

Why Does Test Research Have Limited Relevance to Practice?

STUDY QUESTION AND OBJECT OF STUDY

The first reason that test research has limited relevance to practice is the nature of the questions that are usually addressed. The practical utility of estimation of sensitivity, specificity, and LR for a particular test in the diagnosis of a particular disease is not always obvious *(7, 13)*. Consider, for example, the diagnostic workup for patients suspected of deep vein thrombosis (DVT). The relevant research question for patients suspected of DVT would be: "Given patient history and physical examination, which subsequent tests (e.g., d-dimer measurement) truly provide added information to predict the presence or absence of DVT?" The probability of disease presence and quantifying which tests independently contribute to the estimation of this probability should be the objects of study. However, in this respect many studies have aimed only to estimate the sensitivity and specificity of the d-dimer assay. When this is the object of a study, it is only the probability of obtaining a positive or negative test result that is addressed, rather than the probability of disease presence. Moreover, the focus is on the value of a single test rather than on the value of that test in combination with other, previous tests, including patient history and physical examination. We may say that the object of research is the test rather than the (probability of) disease. Hence the term test research.

TEST CHARACTERISTICS ARE NOT FIXED

The second reason that results from test research have limited relevance is that a test's sensitivity, specificity, LR, and ROC area tend to be taken as properties or characteristics of a test. This, however, is a misconception, as we discussed recently *(13)*. It is widely accepted that the predictive values of a test vary across patient populations. However, several studies have empirically shown that the sensitivity, specificity, and LR of a test may vary markedly, not only across patient populations *(14)* but also within a particular study population *(13, 15–17)*. Within different patient subgroups, defined by patient characteristics or other test results, a particular test may have different sensitivities and specificities. This is because all diagnostic results obtained from patient history, physical examination, and additional tests are to some extent related to the same underlying disorder. For example, immobility, gender, and use of oral contraceptives are associated with the development of, and thus the presence of, DVT. In turn, the presence of DVT determines the presence of symptoms and signs and also (the probability of finding) a positive d-dimer assay result. Accordingly,

via the underlying disorder, all diagnostic results are somehow correlated and thus mutually determine each other's sensitivity, specificity, and LR to various extents *(13, 15–17)*. A single value of a test's sensitivity, specificity, LR, ROC area, or predictive value that applies to all patients of a study sample does not exist. Hence, there are no fixed test characteristics.

selection bias

The most widely acknowledged limitation of test research is that studies often apply an improper patient recruitment and study design *(1–3, 7)*. Investigators often select study participants among those who underwent the reference test in routine practice, i.e., selection based on a "true" presence or absence of the disease. The results of the test(s) under study are retrieved from the medical records and then compared across those with and without the disease. Such a case–control design commonly leads to selection bias, known as verification, workup, or referral bias *(9, 18, 19)*.

Although such patient recruitment methods and study designs have decreased in the past decade, test research is still frequently based on individuals selected based on their final diagnosis *(1–3)*. The need for proper patient recruitment is extensively addressed in the STARD checklist *(11, 12)*. Study participants should be selected in agreement with the indication for diagnostic testing in practice, i.e., on their suspicion of having a particular disease, rather than on the presence or absence of that disease. Such unbiased selection of study participants may indeed be problematic for diagnostic laboratories or imaging centers that do not have access to consecutive series of patients suspected of having the disease. Moreover, most hospital databases code patients according to their final diagnosis rather than by their presenting symptoms or signs. The use of a system to register patients not only on their final diagnosis but also on their clinical presentation would enhance the validity and clinical relevance of diagnostic accuracy research *(20)*.

Proposed Approach for Diagnostic Accuracy Research

We believe that to serve practice, the point of departure and the multivariable and probabilistic character of the diagnostic workup should be reflected in the objective, design, analysis, and presentation of studies of diagnostic accuracy. The aim is to relate the probability of disease presence to combinations of test results, following their typical chronology in practice. The predictive accuracy of the initial tests (including patient history and physical examination) should be estimated first, and the added value of more burdening and costly tests should be estimated subsequently. Hence, all tests typically applied in the workup need to be documented in each patient, even if a study focuses on a particular test. Consider again the question whether the d-dimer assay is relevant to the diagnosis of DVT. A consecutive series of patients suspected of DVT should be selected. The history, physical examination, and d-dimer result should be obtained from each patient. Subsequently, each patient "undergoes" the best reference test currently available; in this example, it would be repeated leg ultrasound. What to do in the absence of a single reference test or when it is unethical to perform the reference test in each patient has been described elsewhere *(8, 10, 21, 22)*.

Because the d-dimer assay will always be applied after history taking and physical examination, the statistical analysis requires a comparison of the (average) probability of disease presence without and with the d-dimer assay, overall or in subgroups. Such sequential modeling of the diagnostic probability as a function of different combinations of test results can be done using, e.g., multivariable logistic regression. Such multivariable analyses account for the mutual dependencies between different test results and thus indicate which tests truly do and which do not independently contribute to the estimation of the probability of disease presence. In addition, various orders of diagnostic testing can be analyzed. The result of such analysis is the definition of one or more diagnostic prediction models including only the relevant tests. If needed, such prediction models can be simplified to obtain readily applicable diagnostic decision rules for use in practice. Various authors have applied or described the details of such an analytical approach *(20, 23–27)*.

Multivariable diagnostic prediction models or rules are not the solution to everything. They may have several drawbacks, such as overoptimism, although methods have been described to overcome some of these drawbacks *(23)*. The need for multivariable modeling in diagnostic research, however, is not different from other types of medical research, such as etiologic, prognostic, and therapeutic research. It is not the singular association between a particular exposure or predictor and the outcome that is informative, but their association independent of other factors. For example, in etiologic research, investigators never publish the crude estimate between exposure and outcome only, but always the association in view of other risk factors (confounders), using a multivariable analysis as well *(13)*. Similarly, in diagnostic accuracy research, multivariable modeling is necessary to estimate the value of a particular test in view of other test results. As in other types of research, such knowledge cannot be inferred from singular, univariable test parameters *(7, 8, 13)*.

Fortunately, a multivariable approach in design and analysis aiming to quantify the independent value of diagnostic tests has gained approval *(20, 23–27)*. In addition, the above study question on the added value of the d-dimer assay in diagnosing DVT has been evaluated in such a way. The d-dimer assay appeared to have an added predictive value to patient history and physical examination, particularly in patients who have a low clinical probability of DVT *(27)*.

Clinical Examples

We now present two clinical examples illustrating how results from a single or univariable test approach can mislead.

In an Australian study, 399 consecutive dyspeptic patients referred for endoscopy underwent two tests, the rapid urease test and the ^{13}C breath test, for *Helicobacter pylori* (HP) with endoscopy as the reference test *(28)*. The investigators found large differences in the test results between patients with a normal and abnormal endoscopy. The sensitivity and specificity were 96% and 67% for the rapid urease test and 91% and 82% for the ^{13}C breath test. The authors concluded that the HP tests might have potential for the initial evaluation of dyspepsia and needed further evaluation in general practice. A second study was done by Weijnen et al. *(26)*. Using a sequential multivariable approach, they found in a consecutive series of 565 dyspeptic patients referred for endoscopy that the HP test did not add diagnostic information to the predictors from history (i.e., history of ulcer, pain on empty stomach, and smoking). The ROC area of the model with only predictors from patient history was 0.71, which was increased to only 0.75 ($P = 0.46$) after addition of the HP test result. They concluded that HP testing in all dyspeptic patients has no value in addition to history taking.

Cowie et al. *(29)* studied a consecutive series of 122 patients suspected of heart failure. They measured in each patient the plasma concentrations of three natriuretic peptides, A-type natriuretic peptide (ANP), N-terminal ANP, and B-type natriuretic peptide (BNP), as well as the presence or absence of heart failure, using consensus diagnosis based on chest radiography and echocardiography as the reference test. They found that the mean concentration of each natriuretic peptide separately (single-test approach) was significantly greater in the patients with heart failure (all $P < 0.001$). They also evaluated all three together in a multivariable logistic prediction model. Only the BNP measurement remained significantly associated with heart failure presence, whereas the other two did not add any predictive information.

Both examples show that one may qualify a test differently (commonly more promisingly) when only the results of a univariable or single-test approach are considered. Evaluating a particular test in view of other test results and accounting for mutual dependencies may decrease or even diminish its diagnostic contribution, simply because the information provided by that test is already provided by the other tests. Because in real life any test result is always considered in view of other patient characteristics and test results, diagnostic accuracy studies that address only a particular test and its characteristics have, in our view, limited relevance to practice. Indeed, as shown by Reid et al. *(30)*, test characteristics are hardly ever actually used by practitioners.

Is There a Place for Test Research?

There are two situations in which pure test research, i.e., studies aiming to estimate the diagnostic accuracy indices of a single test, is indicated. The first situation is when a diagnosis is indeed set by only one test and other test results are not considered. This is, in our view, reserved to the context of screening for preclinical stages of a particular disease only: e.g., screening for breast cancer, prostate cancer, or cervical cancer. Such screening may be considered as a specific case of diagnosis, concerned with the early detection of a disease in a particular age and sex group. Here, only the screening test is considered in the diagnostic process; other patient characteristics or test results are commonly not available and therefore cannot modify the sensitivity, specificity, LR, and predictive values of the screening test. Accordingly, these indices, as estimated from a particular study sample, may be considered characteristics or constants for the corresponding source population. In the presence of a positive screening result, patients are commonly referred for further diagnostic workup. Other test results then become involved, and mutual dependencies between the screening test and these other tests start to play a role, demanding a multivariable approach in design and analysis.

The second situation, as suggested previously, is in the initial phase of developing a new test or evaluating an existing test in a new context; single-test evaluations in these circumstances may be useful for efficiency reasons *(4, 6, 7, 25)*. Such initial test research should apply a case–control approach, preferably starting with a sample of patients with the disease (cases) and a sample of healthy controls. If the test cannot differentiate between these two extreme or heterogeneous outcome categories, the test development process would likely be terminated. In such instances, it will be unlikely that the test does show discriminative value in patients suspected of having the disease, i.e., the population for which the test is intended, because these patients present with similar disease profiles, leading to an even more homogeneous case mixture. However, once the test does yield "satisfactory" diagnostic indices in such an initial test research study, we believe that its independent predictive contribution to existing diagnostic information in a clinical context can and must still be quantified by the above proposed approach.

We gratefully acknowledge The Netherlands Organization for Scientific Research for their support (No. 904-66-112).

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.
2. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.

3. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85–91.
4. Fryback D, Thornbury J. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88–94.
5. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–91.
6. van der Schouw YT, Verbeek ALM, Ruijs JHJ. Guidelines for the assessment of new diagnostic tests. Investig Radiol 1995;30:334–40.
7. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539–41.
8. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337–8.
9. Knottnerus JA. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:237pp.
10. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002;222:604–14.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;49:1–6.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.
13. Moons KGM, Harrell FE. Sensitivity and specificity should be deemphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–2.
14. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986;104:66–73.
15. Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64–71.
16. Levy D, Labib SB, Anderson KM, Christiansen JC, Kanell WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990;81:815–20.
17. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.
18. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30.
19. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:297–315.
20. Oostenbrink R, Moons KGM, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003;56:501–6.
21. Moons KGM, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic research. J Clin Epidemiol 2002;55:633–6.
22. Bossuyt PPM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–7.
23. Harrell FE. Regression modeling strategies, 1st ed. New York: Springer-Verlag; 2001:600pp.
24. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488–94.
25. Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology 1999;10:276–81.
26. Weijnen CF, Numans ME, de Wit NJ, Smout AJ, Moons KG, Verheij TJ, et al. Testing for *Helicobacter pylori* in dyspeptic patients suspected of peptic ulcer disease in primary care: cross sectional study. BMJ 2001;323:71–5.
27. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, et al. Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis. Thromb Haemost 1999;81:493–7.
28. Fraser AG, Ali MR, McCullough S, Yeates NJ, Haystead A. Diagnostic tests for *Helicobacter pylori*–can they help select patients for endoscopy? N Z Med J 1996;109:95–8.
29. Cowie MR, Struthers AD, Wood DA, Coats AJ, Thompson SG, Poole-Wilson PA, et al. Value of natriuretic peptides in assessment of patients with possible new heart failure in primary care. Lancet 1997;350:1349–53.
30. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. Am J Med 1998;104:374–80.

Karel G.M. Moons*
Cornee J. Biesheuvel
Diederick E. Grobbee

Julius Center for Health Sciences and Primary Care
University Medical Center
Utrecht, The Netherlands

*Address correspondence to this author at: Julius Center for Health Sciences and Primary Care, University Medical Center, P.O. Box 85500, 3508 GA Utrecht, The Netherlands. Fax 31-30-2505485; e-mail K.G.M.Moons@jc.azu.nl.

DOI: 10.1373/clinchem.2003.024752

Using the Principles of Randomized Controlled Trial Design to Guide Test Evaluation

Sarah J. Lord, MBBS, MS, Les Irwig, MBBCh, PhD, Patrick M. M. Bossuyt, PhD

The decision to use a new test should be based on evidence that it will improve patient outcomes or produce other benefits without adversely affecting patients. In principle, long-term randomized controlled trials (RCTs) of test-plus-treatment strategies offer ideal evidence of the benefits of introducing a new test relative to current best practice. However, long-term RCTs may not always be necessary. The authors advocate using the hypothetical RCT as a conceptual framework to identify what types of comparative evidence are needed for test evaluation. Evaluation begins by stating the major claims for the new test and determining whether it will be used as a replacement, add-on, or triage test to achieve these claims. A flow diagram of this hypothetical RCT is constructed to show the essential design elements, including population, prior tests, new test and existing test strategies, and primary and secondary outcomes. Critical steps in the pathway between testing and patient outcomes, such as differences in test accuracy, changes in treatment, or avoidance of other tests, are displayed for each test strategy. All differences between the tests at these critical steps are identified and prioritized to determine the most important questions for evaluation. Long-term RCTs will not be necessary if it is valid to use other sources of evidence to address these questions. Validity will depend on issues such as the spectrum of patients identified by the old and new test strategies. ***Key words:*** *diagnostic techniques and procedures/standards; sensitivity and specificity; randomized controlled trials as topic; outcome assessment (health care).* ***(Med Decis Making 2009;29:E1–E12)***

Tests are generally used to provide diagnostic, prognostic, or predictive information to guide treatment decisions. Test evaluation is undertaken to investigate whether this information improves patient outcomes or whether the new test produces other benefits, such as improved safety or reduced costs.

Randomized controlled trials (RCTs) that allocate patients to the new test strategy or current best practice provide ideal evidence of the net benefits or harms of introducing a new test. These RCTs should have long-term follow-up to capture all immediate and downstream consequences of testing, including the effects of any changes in treatment. However, long-term RCTs comparing test-plus-treatment strategies may not always be available, feasible, or even necessary. In some situations, more efficient RCT designs may be possible.[1,2] In other situations, comparative evidence about the safety and accuracy of the test from observational studies may suffice because trials have already demonstrated the benefits of treatment for the cases detected.[3]

Received 11 September 2008 from NHMRC Clinical Trials Centre (SJL) and Screening and Test Evaluation Program, School of Public Health (SJL, LI), University of Sydney, Sydney, Australia, and Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, the Netherlands (PMMB). Revision accepted for publication 17 March 2009.

This article is part of the White Paper series from the Agency for Health Care Research and Quality (AHRQ) Effective Health Care Program. An earlier version of this article was presented at the Diagnostic Test Evaluation Working Meeting, Rockville, MD, May 28–29, 2008.

See also the following related articles:

AHRQ Effective Health Care Program White Paper Series

Note From Editor 634

Proposals for a Phased Evaluation of Medical Tests E13

Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests: Uses and Limitations E22

Additional Patient Outcomes and Pathways in Evaluations of Testing E30

DOI: 10.1177/0272989X09340584

Address correspondence to Sarah J. Lord, MBBS, MS, National Health and Medical Research Council Clinical Trials Centre, University of Sydney, Level 6, Medical Foundation Building, 92-94 Parramatta Rd, Locked Bag 77, Camperdown, NSW 2050, Australia; e-mail: slord@ctc.usyd.edu.au.

Optimizing the efficiency of an RCT or determining what other study designs may suffice begins by describing the pathway by which the new test is expected to improve patient outcomes. Careful scrutiny of this pathway is undertaken to identify critical steps that will determine the effectiveness of the test and the comparisons needed to investigate these steps. For instance, if a new test is intended to reduce patient morbidity by detecting additional cases of disease, the most important question is as follows: What is the effect of treatment in the extra cases detected? An efficient RCT design would therefore be to focus on comparing treatment v. no treatment in this patient subgroup only: those who test negative on the existing test but positive on the new test.

The same principles apply when planning a systematic review of evidence for test evaluation. If the benefits of an existing test-plus-treatment strategy have already been established, the type of evidence needed depends on how it will be used to alter the existing pathway. This is determined by the proposed attributes of the new test and whether it will be positioned as a replacement for the existing test, an add-on test after the existing test, or as a triage test before the existing test.[4]

To date, guidance for systematic reviews of diagnostic tests has predominantly focused on the methods for assessing test accuracy rather than how to assess the consequences of test results and other test attributes on patient outcomes. The purpose of this article is to 1) describe how a hypothetical RCT offers a useful conceptual framework to identify what types of comparative evidence are needed to evaluate a new test; 2) describe how the type of evidence needed varies according to whether the new test will be used as a replacement, add-on, or triage test and the intended benefits; and 3) identify situations where RCTs assessing the entire test-plus-treatment pathway are essential for conclusions about the impact of a new test on patient outcomes or where studies assessing test accuracy, safety, and other immediate or intermediate outcomes may suffice.

USING THE RCT ANALOGY IN TEST EVALUATION

Planning an evaluation of a new test is analogous to developing a protocol for an RCT for the same purpose. The first task is to clarify the claim and list the primary intended changes in patient, cost, or other health service outcomes. This is analogous to defining the primary study objective for an RCT. All other potential changes in outcomes can then be listed as secondary objectives.

Potential patient benefits of testing include reducing patient mortality or morbidity and improving health-related quality of life or other effects, such as reduced patient discomfort, anxiety, or inconvenience. To define all potential changes in outcomes, clinical experts involved in the diagnosis and management of the target test population can advise on the most likely role of the new test relative to current practice and consider ''How will patients be better off?'' ''How might they be worse off?'' It may be helpful to refer to a checklist of possible patient outcomes to assist this process (see the article by Bossuyt and McCaffery in this issue[5]).

The incremental benefits and harms of a new test may occur along one or more pathways. One pathway involves a series of steps linking improved test accuracy with changes in patient management and the effects of these management changes on patient outcomes. Changes in patient outcomes can occur along this ''test-treatment'' pathway because of a change in treatment or further testing. In another pathway, a new test may affect patient outcomes through other attributes of the procedure itself, such as safety and acceptability, without changing patient management. Tests may also affect patient outcomes through the patients' emotional, cognitive, or behavioral response to their test result and clinical management.[5]

The next task is to determine what type of evidence is needed to investigate these claims. We propose that those evaluating evidence for decision making imagine the design of an RCT to measure all specified patient outcomes and construct a flow diagram to map out the key elements of this trial, including the target test population, prior tests, index test strategy, comparator strategy, and outcomes as it would appear in an RCT protocol. This flow diagram can be used to determine what type of comparative evidence is needed to demonstrate a difference in patient outcomes between the new v. existing test strategy. The validity of the evidence available can be appraised using this hypothetical RCT as a benchmark. Conceptually, this approach is fundamentally different to and should precede the use of a decision model to integrate the evidence available or the development of a clinical algorithm to guide practice, although the flow diagram can be adapted for these purposes should the test evaluation identify adequate comparative evidence.

CONSTRUCTING A TEST EVALUATION FLOW DIAGRAM

Display Existing Test Strategy

The first step when constructing the flow diagram is to display the target population and the current best test-treatment pathway for managing these patients. This pathway represents the comparator strategy and may involve no prior testing if the new test is intended for primary screening or for monitoring treatment, or it may be an existing test strategy with subsequent management defined by the test result.

Display New Test Strategy

The next step is to define the new test and describe the alternative test-treatment pathway proposed using this test. This involves identifying where in the existing sequence of tests the new test will be used; whether it will be used as a replacement, add-on, or triage test for existing tests; and management following positive and negative test results. This pathway should be displayed alongside the existing test pathway on the flow diagram. Prior tests that are common to each pathway can be listed with the definition of the target population. Figure 1a shows a generic test evaluation flow diagram that can be used for a replacement test. Figure 1b,c shows how the test-treatment pathway varies if the new test will be used as an add-on or triage test.

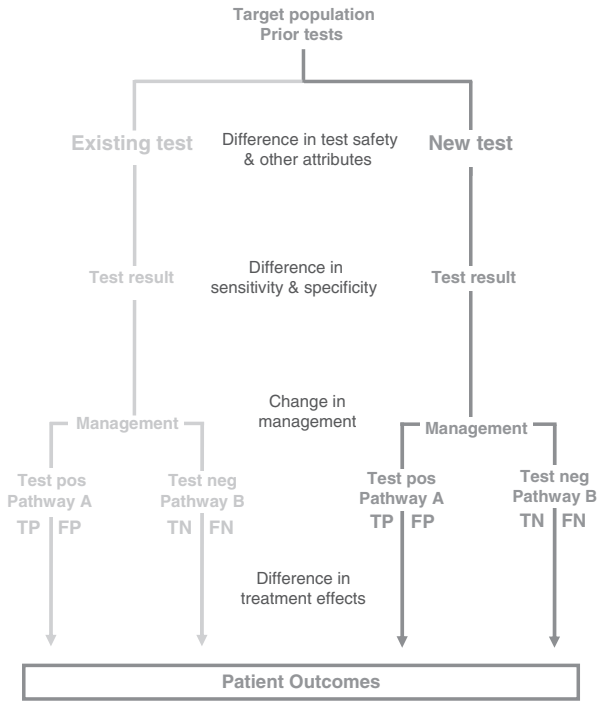
Identify Critical Comparisons

All differences between the new and existing test strategy that can be linked to immediate or downstream consequences for the patient, both benefits and harms, can be labeled as critical comparisons on the flow diagram. These differences will determine the effectiveness of the new test and can be used to formulate the research questions for evaluation. As the signposted in Figure 1, differences in test safety and test accuracy always deserve consideration. If a difference in test sensitivity or specificity is identified, changes in management for the extra true- or false-positive or negative test results also need to be considered together with evidence about the impact of these changes in management on patient outcomes.

Differences in test attributes along other pathways will be at least as important in many cases.[5] These may include:

- other consequences of the test procedure itself, such as improved access for patients;

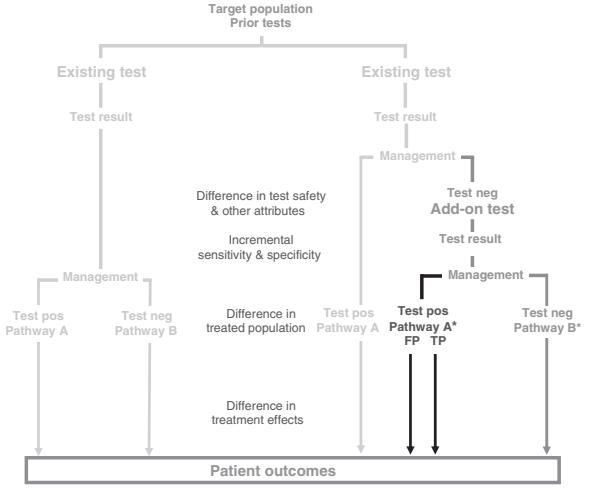
a. The replacement test



TP = true positive, FP = false positive, TN = true negative, FN = false negative

b. The add-on test

Difference in test-treatment pathway using add-on test shown in black



TP = true positive, FP = false positive
Pathway A* includes patients testing positive on the add-on test but negative on the existing test who would not have been assigned to treatment A using the existing test strategy.

(continued)

(continued)

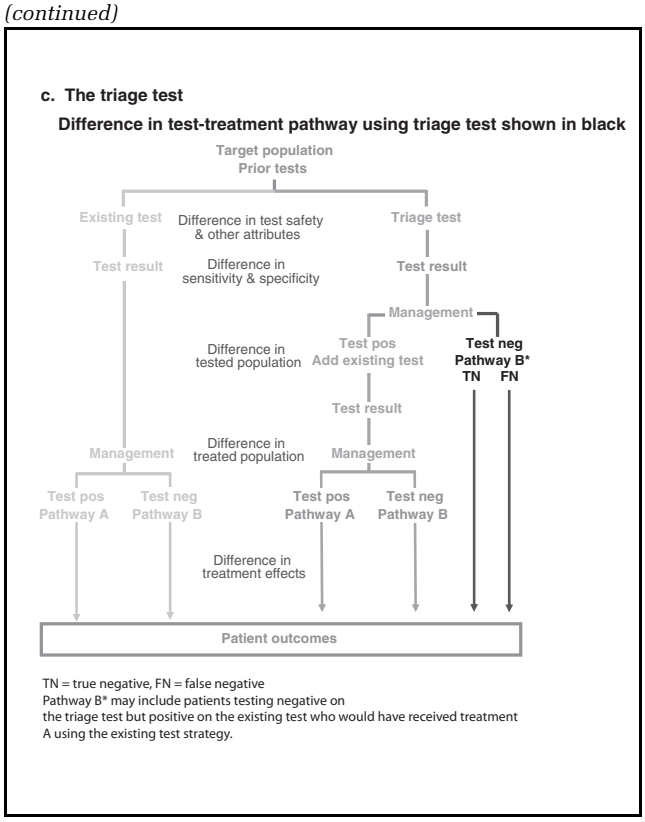


Figure 1 Test evaluation flow diagrams.

- other consequences of the test results, such as additional clinical information about prognosis without altering treatment selection; or
- other consequences of clinical decisions, such as increased adherence to treatment or adoption of healthy behaviors among patients testing positive using the new test.

Priority can be assigned to pathways and comparisons within pathways that are directly associated with higher order effects such as patient mortality and morbidity. For example, if a new test is more sensitive than the existing test, priority is assigned to the test-treatment pathway. Within this pathway, the difference in treatment effects for the extra cases detected is the most critical comparison determining changes in patient outcomes.

The key elements of the hypothetical RCT can be used to define criteria for selecting relevant evidence to address each comparison. Multiple flow diagrams may be needed if the role of the new test or the type of comparator differs for different patient groups—for example, if the test will be used as an add-on test in primary care but as a triage test in tertiary care. Three examples to show how the flow diagram can be used to map out critical comparisons, select evidence, and judge the need for long-term RCT are discussed below and summarized in Table 1.

IDENTIFYING WHAT TYPE OF COMPARATIVE EVIDENCE IS NEEDED FOR TEST EVALUATION

The Replacement Test

A new replacement test may be introduced to improve patient outcomes by improving treatment selection along the test-treatment pathway if it is more sensitive and/or specific than the existing test or by providing benefits along other pathways due to other attributes.

Critical Comparisons and Evidence

1) If the new test is intended to be more sensitive than the old test, the critical comparisons are as follows: First, does treatment of the extra true-positive cases detected improve patient outcomes? Second, what is the difference in sensitivity and specificity of the new test for detecting extra cases of disease? Trials may have already demonstrated the efficacy of treatment among cases detected by the existing test to help address the first question. However, this evidence may not apply to the extra cases detected by the new test if they represent a different spectrum of disease, as discussed below for add-on tests.

2) If the major intended benefit of the new test is improved specificity, the critical comparisons are as follows: First, what are the benefits of fewer false-positive findings? Second, what is the difference in sensitivity and specificity of the new test? If the new test is more specific but less sensitive than the old test, careful assessment of the tradeoff between these benefits v. the harms of additional false negatives will be needed.

3) Alternatively, if the new test is intended to provide other attributes such as improved safety, the critical comparisons are as follows: First, is the new test at least as sensitive and specific as the existing test so patient management is not compromised? Second, what is the difference in adverse event rates or other relevant outcomes?

Example: Liquid-Based Cytology as a Replacement for the Pap Test in Cervical Screening

Using the flow diagram in Figure 1a, consider the critical comparisons and types of evidence needed to evaluate liquid-based cytology (LBC) as

Table 1 Examples of Comparative Evidence Needed to Assess the Impact of a New Test on Patient Outcomes and Optimal Study Designs

1. Replacement test

Liquid-based cytology (LBC) as a replacement for Pap tests for detection of precancerous cervical lesions in screening populations
Intended benefits: reduced unsatisfactory slide rate leading to reduced patient recall, improved convenience, and adherence to screening protocol

Comparison	Rates and Consequences[a]	Optimal Study Design
1. Test-treatment pathway		
i. Difference in test sensitivity and specificity		Comparative accuracy study
True positives	No difference[7,8]	RCT *or* cross-sectional study
True negatives	No difference[7,8]	with pairwise comparison
False positives	No difference	Complete verification *or* verification
False negatives	No difference	of discordant test results to compare relative TP and FP rates
Consequences of test results: No change in test-treatment pathway		
ii. No difference in treated population		
iii. No difference in treatment effects		
Other pathways		
2. Test safety	No difference in adverse events of test procedure	NA
3. Unsatisfactory slide rate	Fifteen fewer unsatisfactory slides per 1000 tested[9]	Short-term RCT: new v. existing test Compare unsatisfactory slide rate Medium-term RCT: new v. existing test Compare adherence ≥ 2 screening rounds
Immediate consequences:		
Assess reduced recall rate		
Assess improved patient convenience		
Potential medium-term consequences:		
Assess adherence to screening protocol		

Long-term RCT of test-plus-treatment strategy not required

2. Add-on test

Addition of breast MRI for the detection of multiple tumor foci in women with early breast cancer planned for BCS following a negative finding on mammogram and ultrasound (existing test strategy)
Intended benefits: detection of extra cases of multicentric or multifocal disease that will benefit from conversion from BCS to mastectomy

Comparison	Rates and Consequences	Optimal Study Design
1. Test-treatment pathway		
i. Difference in test sensitivity and specificity		
True positives	+91 per 1000 tested[14] change in treatment	Comparative accuracy study RCT *or* cross-sectional study
True negatives	Not verified	with pairwise comparison
False positives	+49 per 1000 tested[14] adverse events, anxiety	Complete verification *or* verification of new test-positive results to compare TP:FP ratio
False negatives	Not verified	
Consequences of test results: Change in test-treatment pathway		
ii. Difference in treated population		Short-term RCT: new v. existing test strategy
Single-arm studies indicate that an additional 80 per 1000 women tested convert from BCS to mastectomy[14]		Compare mastectomy rates

(continued)

Table 1 (continued)

iii. Treatment effects of converting to mastectomy	Long-term RCT: mastectomy v. BCS + XRT
Evidence is available for cases defined by mammography and ultrasound:	Population: MRI-positive/mammography and ultrasound-negative cases
RCT evidence of similar treatment effects using BCS + XRT or mastectomy for unifocal disease (test negative)[10]	Compare survival, recurrence-free survival. Morbidity, health-related quality-of-life
Observational evidence of worse prognosis for multifocal (test-positive) v. unifocal (test-negative) disease[11]	
This evidence may not apply to the extra cases detected by MRI	
iv. Consequences of extra false positives	
Compare types, rates, and adverse events of unnecessary test and treatments	Prospective consecutive test series to estimate frequency, type, and outcomes of unnecessary tests and treatment
Assess differences in patient anxiety	Short-term RCT: new v. existing test strategy Compare patient anxiety
Other pathways	
2. Test safety	Prospective consecutive test series to estimate frequency of MRI adverse events
No additional adverse events of test procedure if MRI contraindications observed	
RCT of treatment required to compare effects of converting to mastectomy in extra true positives	

3. Triage test

D-dimer for triage of low-risk patients with suspected deep venous thrombosis who would otherwise be referred to ultrasound

Intended benefits: to improve patient convenience and costs by excluding deep venous thrombosis and avoiding ultrasound in patients with a negative D-dimer test

Comparison	Rates and Consequences	Optimal Study Design
1. Test-treatment pathway		
i. Difference in sensitivity and specificity		
True positives	70 per 1000 tested[16] receive ultrasound	Comparative accuracy study RCT *or* cross-sectional study with pairwise comparison Complete verification *or* verification of discordant test results to compare TN:FN ratio
True negatives	581 per 1000 tested[16] avoid ultrasound	
False positives	+ 346 per 1000 tested[16] receive ultrasound	
False negatives	+ 2 per 1000 tested[16] delayed diagnosis	
Consequences of true negatives: Change in test-treatment pathway		
ii. Difference in tested population		
Assume all true negatives avoid ultrasound		
Assess effects of avoiding ultrasound: convenience accessibility		Short-term RCT: new v. existing test strategy Compare patient acceptability, accessibility, utilities
iii. Consequences of extra false negatives		
Difference in treated population: assume all false negatives experience delay in detection and treatment		Long-term RCT of entire test-plus-treat strategy to compare survival, morbidity
Assess effects of early v. delayed treatment		*If infeasible:* Short-term RCT: new v. existing test Compare time to diagnosis and treatment Epidemiological evidence about the natural history of DVT to estimate risk of morbidity and mortality

(continued)

Table 1 (continued)

Other pathways		
2. Test safety	No difference in adverse events of test procedure	NA
3. Potential role as add-on test		Short-term RCT: new v. existing test
The availability of a simple D-dimer could lower the threshold for testing: more patients undergo D-dimer testing than would otherwise be referred for ultrasound based on clinical assessment: more true and false positives.		Compare test-ordering practices
Evidence from comparative accuracy studies and short-term RCT may suffice		

BCS, breast-conserving surgery; DVT, deep vein thrombosis; FP, false positive; MRI, magnetic resonance imaging; NA, not applicable; RCT, randomized controlled trial; TP, true positive; XRT, radiotherapy.
a. These rate estimates are based on data extracted from cited publications for the purpose of illustration only.

a replacement for the conventional Pap test for cervical cancer screening (Table 1). We already have evidence that Pap test screening programs reduce the incidence and mortality of cervical cancer.[6] For the purpose of this simplified example, we will assume the major claims for LBC are reduced cervical cancer incidence due to improved test sensitivity and reduced unsatisfactory slide rates.

Test-Treatment Pathway

The first priority is to compare the sensitivity and the specificity of LBC v. the Pap test. Two meta-analyses have reported that LBC has a similar sensitivity and specificity as the Pap test for the detection of high-grade cervical abnormalities.[7,8] On the basis of this evidence, we conclude that LBC will not directly lead to a change in management. A long-term RCT of the entire new test-plus-treatment strategy is not needed because the efficacy of the existing test-plus-treatment strategy has already been established, and the new test will not alter this pathway.

The exception is if the 2 tests have similar sensitivity and specificity but do not have perfect agreement and detect patients in a different spectrum of disease. An example would be comparing 2 cytology tests using the total number of true low-grade and high-grade lesions detected, where one test detects a higher proportion of high-grade cytological abnormalities and a lower proportion of low-grade cytological abnormalities than the other test.

Other Pathways

The next step is to assess differences in test safety and other attributes of the test that may be linked to patient outcomes along other pathways. LBC is a safe procedure and is associated with the same level of patient discomfort as the Pap test, so no differences in immediate patient outcomes are anticipated as a result of the test procedure. A comparison of unsatisfactory slide rates using LBC v. Pap tests takes next priority. A reduced unsatisfactory slide rate for LBC may provide immediate patient benefits by reducing patient recall rates, associated inconvenience and costs, and potential downstream benefits, such as improved adherence to screening protocols.

Short-term RCTs would provide ideal evidence about a difference in unsatisfactory slides rates between the 2 test strategies. At least one such trial has been performed.[9] If follow-up was undertaken to the next screening round, such RCTs would also provide valuable evidence about a difference in patient adherence.

Long-term RCTs are not required if assumptions linking these intermediate outcomes to long-term patient benefits, including reduced cancer incidence and mortality, appear to be reasonable.

The Add-On Test

Add-on tests are generally introduced to improve patient outcomes through improved treatment selection by increasing the sensitivity or specificity of a testing strategy. They may be used on all patients, in which case the test-treatment pathway resembles that shown in Figure 1a for replacement tests, or reserved for a subset of patients—for example, the addition of a more sensitive test for those testing negative on the existing test, as shown in Figure 1b and discussed in the breast magnetic resonance imaging (MRI) example below.

Critical Comparisons

The critical comparisons for add-on tests are first to establish the treatment effects of detecting extra cases of disease if adding the new test increases sensitivity or the benefits of avoiding further tests or treatment if adding the new test increases specificity. Second, the incremental sensitivity and specificity of adding the new test need to be assessed to estimate the additional proportion of patients tested who will benefit.

Traditional cross-sectional paired accuracy studies to compare the sensitivity and specificity of the new v. existing test strategy in all patients may not be required. As highlighted in Figure 1b, the only difference in management between the new and existing test strategies occurs among patients testing positive using the new test following a negative result on the existing test. Thus, verification of test results by the reference standard for this subpopulation with discordant test results is sufficient. All other patients would receive the same treatment in each arm of the hypothetical RCT and do not contribute beyond chance to any difference in treatment outcomes between the tests.

Example: MRI as an Add-On Test for Staging Early Breast Cancer

Breast MRI is proposed to detect additional tumor foci in women with a diagnosis of early breast cancer on mammography and ultrasonography planned for breast-conserving surgery (BCS; Table 1). RCTs have demonstrated that BCS plus adjuvant radiotherapy is a safe and effective alternative to mastectomy with a similar risk of disease recurrence in women with stage I to II disease.[10] Mastectomy is recommended if mammography detects multicentric or multifocal disease based on evidence that this population is at higher risk of local recurrence following BCS plus radiotherapy than women with unifocal disease.[11,12] The major intended benefit of MRI is to improve overall and/or recurrence-free survival by detecting extra cases with multicentric or multifocal disease who will benefit from conversion from BCS to mastectomy. This treatment comparison therefore takes priority for the evaluation. The other critical comparisons are the magnitude of the increase in sensitivity, the extent to which true-positive test findings lead to a change in management and therefore patient outcomes, and the consequences of false-positive findings. These issues are discussed separately below.

Test-Treatment Pathway

Randomized comparisons are needed to assess the efficacy of converting from BCS to mastectomy in women with additional tumor foci detected by MRI that are mammography-occult (pathway A*, Figure 1b). Should the evaluation proceed without this RCT evidence? This depends on judgments about the plausibility of assumptions that the treatment effects observed in women with mammogram-detected disease will equally apply to this new subpopulation, as well as the potential consequences should these assumptions later be proven incorrect.

Table 2 uses a hypothetical example to describe how estimates of treatment effects for the extra cases detected by a new test may vary according to different assumptions about patient prognosis and treatment response. In theory, the absolute benefits of a new treatment depend on 3 factors: the patient risk of future disease events without this treatment (prognosis), the relative effectiveness of the new treatment, and the risks of treatment. The absolute risk reduction equals the patient baseline risk times the relative risk reduction minus the risks of treatment.

Applying these concepts to the breast MRI example, if it is reasonable to assume mastectomy provides the same relative effects for patients detected by either test, the number of tumor recurrences avoided at 10 years per 1000 extra cases detected will vary proportionally to the prognosis of these cases when treated by standard BCS plus adjuvant radiotherapy alone. On one hand, if the extra cases show a prognosis similar to cases detected by mammography and ultrasonography, the same absolute treatment benefits can be expected (scenario 1, Table 2). One the other hand, if the extra cases show a similar prognosis to mammography and ultrasonography negative ''noncases,'' the addition of MRI will not be warranted based on existing RCT evidence that BCS is adequate for this low-risk group (scenario 2, Table 2).

Alternatively, patient prognosis and treatment response for the extra MRI-detected cases may lie somewhere between these extremes (scenarios 3 and 4, Table 2). Observational studies can sometimes offer useful evidence about differences in prognosis between patient groups. However, prognostic studies are unlikely to be feasible in this example where the goal would be to compare long-term outcomes for the extra cases of multicentric or multifocal disease detected by MRI with cases detected by mammography alone when both groups are managed with BCS plus adjuvant radiotherapy. Even so, regardless of whether prognostic information is available, RCTs would still be needed to test the assumption that mastectomy provides the same relative effects for patients detected by either test. This

Table 2 Estimating the Effects of Finding Extra Cases

Hypothetical Example: Add-On Test Detects Extra Cases of Disease	Risk of Disease Events, %					Treatment effects for Extra Cases: Estimated Reduction in Absolute Risk of Disease Events (ARR)[d] for Treatment v. Standard Care
	Existing Test Positive[a]		Existing Test Negative[b]	Existing Test Negative/ New Test Positive[c]		
	Standard Care	Treat	Standard Care	Standard Care	Treat	
Scenario 1: extra cases have same prognosis as cases detected by existing test - assume same relative treatment response	30	10	10	30	10	Same absolute benefits as cases detected by existing test ARR = 20%
Scenario 2: extra cases have same prognosis as noncases defined by the existing test	30	10	10	10	—	Prognosis does not justify a change in treatment
Scenario 3: extra cases have better prognosis than cases detected by existing test - assume same relative treatment response	30	10	10	20	7	Absolute benefits less than for cases detected by existing test ARR = 14%
Scenario 4: extra cases have better prognosis than existing test - assume reduced treatment response	30	10	10	20	10	Absolute benefits less than for cases detected by existing test ARR = 10%

a. Data from hypothetical randomized controlled trial comparing treatment v. standard care for patients detected by existing test.
b. Data from observational study reporting on prognosis for noncases defined by existing test.
c. Estimated risk of disease events in extra cases detected by new test under different assumptions about prognosis and treatment responsiveness.
d. ARR = baseline risk using standard care (prognosis) × relative risk reduction from treatment.

would involve randomizing women with multicentric or multifocal disease to mastectomy or BCS plus adjuvant radiotherapy and comparing the treatment effects between subgroups of women defined by mammography or MRI alone using a statistical test for interaction.

Test Accuracy

If comparative accuracy studies with verification of all test results are not available, cross-sectional studies that verify MRI-positive, mammography-negative patients will suffice to estimate and compare the rate of extra true-positive and false-positive findings. It is not essential to verify concordant negative test results or mammography and ultrasonography test-positive results because this information will not affect treatment decisions or patient outcomes.

Change in Management

Evidence about the impact of the new test on changes in management is needed if there is uncertainty about whether all additional patients testing positive using the new test will receive the same treatment as cases detected by the existing test or whether all patients missed by the existing test will not otherwise receive this treatment. Such uncertainty may arise if further testing occurs before treatment or if other clinical factors or patient preferences influence treatment decisions. Even taking into account the use of needle biopsy to detect false-positive MRI results, some women with a positive MRI finding of multiple tumor foci may still opt for BCS and adjuvant radiotherapy, or some women with a negative mammogram may still proceed to mastectomy due to clinical findings indicating more widespread disease at surgery. Therefore, it will not be possible to infer the additional rate of conversion to mastectomy following MRI based on the rate of extra true-positive MRI-detected cases reported by accuracy studies alone. Ideally, short-term RCTs would be available to quantify the difference in mastectomy conversion rates for true-positive cases with and without MRI. Other sources of evidence would be accuracy studies with a period of follow-up or before-and-after

studies reporting the proportion of women with a true-positive finding who go on to convert from BCS to mastectomy.[13]

Consequences of False-Positive Findings

Finally, what are the consequences of false-positive findings? The immediate consequences, such as unnecessary needle biopsies, surgery, or other interventions and costs, can be determined using data from consecutive series of tested patients that include a period of clinical follow-up. For example, a systematic review of breast MRI included data from accuracy studies that reported on the management of positive MRI findings.[14] This review found that although many false-positive findings are investigated by needle biopsy, conversion from wide local excision to more extensive surgery due to false-positive findings occurred in around 5% of women.

RCTs comparing the new v. existing test strategy would provide the most valid evidence to assess differences in rates of patient anxiety and other adverse events due to these unnecessary interventions. If such trials are not available, conclusions about the benefits of adding MRI have to depend on judgments weighing up the rates and consequences of extra true-positive findings against the rates and consequences of extra false-positive findings (Table 1).

The Triage Test

Triage tests are generally introduced to increase the safety or efficiency of a testing strategy, for example, through the avoidance of more invasive, time-consuming, or costly tests.[4] They present different comparisons to replacement tests because only a proportion of all patients tested avoid the existing test—those testing negative on the triage test, as shown in Figure 1c.

Critical Comparisons

Triage tests often present tradeoffs between the benefits of safer or earlier exclusion of patients without the target condition and the harms of false negatives. The critical comparisons are commonly: is the new test at least as sensitive and specific as the existing test, and what is the difference in adverse event rates or other test attributes beyond accuracy?

Example: D-Dimer as a Triage Test in Suspected Deep Venous Thrombosis

Consider the evaluation of rapid point-of-care D-dimer as a triage test prior to ultrasound in patients with suspected deep venous thrombosis (DVT) clinically assessed as low risk (Table 1). The potential benefits of this strategy include improved patient access and convenience with reduced time and costs due to the avoidance of ultrasound in patients with a negative D-dimer test.[15,16] The potential harms include increased patient morbidity due to missed diagnoses if D-dimer is less sensitive than ultrasound; increased inconvenience, time to definitive treatment, and costs for patients with a positive D-dimer who need to proceed to ultrasound; and/or increased patient anxiety and costs for false-positive D-dimer findings.

Test-Treatment Pathway

The optimal comparative accuracy study would verify all test results with the reference standard. However, the only difference in the use of ultrasound between the test strategies occurs among patients testing negative using D-dimer (pathway B*, Figure 1c). Thus, studies that only compare negative D-dimer results with ultrasound can provide all the required information about clinically meaningful differences in test accuracy for the detection of DVT.

The other key critical comparisons—the effects of delayed treatment for patients with false-negative results—can also be identified for consideration using the flow diagram. The consequences of delayed treatment of DVT are potentially serious and ideally measured by RCTs.

Other Pathways

It is very well possible that the availability of D-dimer lowers the threshold for testing in patients with suspected DVT. This could be explored in short-term RCTs, which would also be ideal to compare patient convenience and other attributes.

When do we need long-term RCTs to assess tradeoffs between the benefits and harms of a new test? Even if feasible, the hypothetical RCT would be unnecessary if there are good comparative studies assessing all critical comparisons, provided assumptions linking this evidence with changes in patient outcomes were judged to be reasonable. If so, decision modeling could be undertaken to integrate these data and quantify differences in patient outcomes.[17]

DISCUSSION

We propose that considering the hypothetical RCT provides a sound conceptual framework for selecting and interpreting evidence to compare

patient outcomes using a new test with current best practice. Using the RCT analogy draws the focus of test evaluation on the most critical comparisons driving the intended changes in patient outcomes. The development of a flow diagram to illustrate the proposed role of the test is helpful to identify the best measure of comparative accuracy, characterize other critical comparisons, and decide whether and what sort of further research is required.

The GRADE working group has recently emphasized the need to assess the consequences of test results on patient outcomes when making recommendations about the quality of evidence for a new test.[18] However, there is little guidance available about what type of evidence is needed to assess these outcomes and how to assess other attributes of the test that may have an impact on patient outcomes.

The USPSTF analytic framework for screening tests provides valuable guidance for mapping out a causal pathway linking testing with patient outcomes and identifying critical steps along the pathway, referred to as ''linkages,'' for investigation, such as test safety, accuracy, and treatment effectiveness.[19] Differences between the new test and existing test strategies at these critical linkages will drive differences in patient outcomes. We provided examples to illustrate that these differences and the types of evidence needed vary according to whether the new test will be used as a replacement, add-on, or triage test and its intended benefits.

RCTs are needed to assess the efficacy of an existing treatment for the extra cases detected by a new, more sensitive test, just as they are needed to assess the efficacy of new treatments. The same applies in reverse if the new test reclassifies some patients with a positive finding on the existing test as ''disease free'' or ''low risk,'' but existing treatment trials have included this patient group. This could happen with a new triage test intended to guide the avoidance of other tests or treatment. An example is the MINDACT trial, which is designed to assess the value of a prognostic gene signature test to identify low-risk women with early breast cancer who can safely avoid chemotherapy.[20]

In situations where comparative evidence of the diagnostic accuracy of a test is not available—for example, where the perfect reference standard does not exist—the flow diagram can be used to identify the most efficient randomized comparisons as an alternative to an RCT of the entire test-plus-treatment strategy. This approach can also be used if the test is intended to guide treatment by providing information to classify patient prognosis or to predict treatment response. In these situations, RCTs that allow a comparison of treatment effects between patients with different test results will also provide optimal evidence. For example, the benefit of testing for estrogen receptor status in women with breast cancer is supported by RCTs of tamoxifen demonstrating an interaction between a patient's estrogen receptor status and response to tamoxifen.

A new test can change patient outcomes by more than one mechanism, producing different effects, in different directions, at different time points in clinical care. All potential consequences of the new test strategy deserve attention when considering the overall effects of testing. The RCT analogy can guide prioritization according to the potential for higher order effects. For example, full-body computed tomography (CT) might lead to reduced morbidity and mortality due to earlier detection of treatable disease in some cases, but these benefits come at a price: the risk of radiation-induced cancer, adverse events of further tests, treatment of spurious findings, and patient anxiety. An evaluation of full-body CT should give priority to identifying evidence about rates of test effects that are associated with changes in patient risk of mortality or serious morbidity. This process can be even more challenging when comparing 2 tests with different attributes, different adverse event profiles, and different sensitivity and specificity.

The construction of a flow diagram based on the hypothetical RCT should not be confused with the process used to construct a decision-analytic model. Decision models represent a later step in evaluation process, where the intended outcomes, causal pathway, and best available evidence have already been defined, although the development of the flow diagram should inform the development of a subsequent decision-analytic model.

Linking evidence from different studies conducted in different populations can never provide evidence about the impact of a new test on patient outcomes of the same strength and quality as an RCT, which captures the entire causal pathway, including the unexpected and unknown pathways. Evaluators must interpret linked evidence with caution. Identifiable uncertainties about effect estimates for critical comparisons and assumptions about linkages in the pathway can be explored using decision modeling,[17] but modeling itself may also be prone to oversimplification and potential bias.

Finally, evaluating tests is not making a choice between using evidence from accuracy studies and

developing RCTs. It should always involve scrutiny of the clinical situation, identification of the health claims of the new tests, and a clear definition of the evidence needed to support or falsify these claims. Regardless of the feasibility of a long-term RCT, both practical and ethical, to quantify all the benefits and harms of testing, the principles of RCT design should guide the identification and interpretation of relevant comparative evidence.

ACKNOWLEDGMENTS

This work was supported in part through an Australian National Health and Medical Research Council Program Grant, no. 402764.

REFERENCES

1. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet. 2000;356:1844–7.

2. Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol. 2009;62: 364–373.

3. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850–5.

4. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089–92.

5. Bossuyt PMM, McCaffery K. Evaluating patient outcomes after testing. Med Decis Making. In press.

6. Peto J, Gilham C, Fletcher O, Matthews FE. The cervical cancer epidemic that screening has prevented in the UK. Lancet. 2004; 364:249–56.

7. Arbyn M, Bergeron C. Liquid compared with conventional cervical cytology: a systematic review and meta-analysis. Obstet Gynecol. 2008;111:167–77.

8. Davey E, Barratt A, Irwig L, et al. Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid-based versus conventional cervical cytology: a systematic review. Lancet. 2006;367:122–32.

9. Ronco G, Cuzick J, Pierotti P, et al. Accuracy of liquid based versus conventional cytology: overall results of new technologies for cervical cancer screening: randomised controlled trial. BMJ. 2007;335:28.

10. Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery on early breast cancer: an overview of the randomised trials. N Engl J Med. 1995;333:1444–55.

11. Leopold KA, Recht A, Schnitt SJ, et al. Results of conservative surgery and radiation therapy for multiple synchronous cancers of one breast. Int J Radiat Oncol Biol Phys. 1989;16: 11–6.

12. Morrow M, Strom EA, Bassett LW, et al. Standard for breast conservation therapy in the management of invasive breast carcinoma. CA Cancer J Clin. 2002;52:277.

13. Guyatt GH, Tugwell PX, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. J Chronic Dis. 1986;39: 295–304.

14. Houssami N, Ciatto S, Macaskill P, et al. Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: systematic review and meta-analysis in detection of multifocal and multicentric cancer. J Clin Oncol. 2008;26: 3248–58.

15. Goodacre S, Sampson F, Stevenson M, et al. Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis. Health Technology Assessment 2006;10(15):1–168.

16. Bates SM, Kearon C, Crowther M, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep venous thrombosis. Ann Intern Med. 2003;138: 787–94.

17. Trikalinos TA, Lau J. Modeling to evaluate benefits and harms of diagnostic tests: uses and limitations. Med Decis Making. In press.

18. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. Br Med J. 2008;336:1106–10.

19. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(suppl):21–34.

20. Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol. 2006;3:540–51.

Annals of Internal Medicine | Academia and Clinic

Systematic Reviews of Diagnostic Test Accuracy

Mariska M.G. Leeflang, PhD; Jonathan J. Deeks, PhD; Constantine Gatsonis, PhD; and Patrick M.M. Bossuyt, PhD, on behalf of the Cochrane Diagnostic Test Accuracy Working Group

More and more systematic reviews of diagnostic test accuracy studies are being published, but they can be methodologically challenging. In this paper, the authors present some of the recent developments in the methodology for conducting systematic reviews of diagnostic test accuracy studies. Restrictive electronic search filters are discouraged, as is the use of summary quality scores. Methods for meta-analysis should take into account the paired nature of the estimates and their dependence on threshold. Authors of these reviews are advised to use the hierarchical summary receiver-operating characteristic or the bivariate model for the data analysis. Challenges that remain are the poor reporting of original diagnostic test accuracy studies and difficulties with the interpretation of the results of diagnostic test accuracy research.

Ann Intern Med. 2008;149:889-897. www.annals.org
For author affiliations, see end of text.

Diagnosis is a critical component of health care, and clinicians, policymakers, and patients routinely face a range of questions regarding diagnostic tests. They want to know whether testing improves outcome; what test to use, purchase, or recommend in practice guidelines; and how to interpret test results. Well-designed diagnostic test accuracy studies can help in making these decisions, provided that they transparently and fully report their participants, tests, methods, and results as facilitated, for example, by the STARD (Standards for Reporting of Diagnostic Accuracy) statement (1). That 25-item checklist was published in many journals and is now adopted by more than 200 scientific journals worldwide.

As in other areas of science, systematic reviews and meta-analysis of accuracy studies can be used to obtain more precise estimates when small studies addressing the same test and patients in the same setting are available. Reviews can also be useful to establish whether and how scientific findings vary by particular subgroups, and may provide summary estimates with a stronger generalizability than estimates from a single study. Systematic reviews may help identify the risk for bias that may be present in the original studies and can be used to address questions that were not directly considered in the primary studies, such as comparisons between tests. The Cochrane Collaboration is the largest international organization preparing, maintaining, and promoting systematic reviews to help people make well-informed decisions about health care (2). The Collaboration decided in 2003 to make preparations for including systematic reviews of diagnostic test accuracy in their Cochrane Database of Systematic Reviews. To enable this, a working group (**Appendix**, available at www.annals.org). was formed to develop methodology, software, and a handbook The first diagnostic test accuracy review was published in the Cochrane Database in October 2008.

In this paper, we review recent methodological developments concerning problem formulation, location of literature, quality assessment, and meta-analysis of diagnostic accuracy studies by using our experience from the work on the Cochrane Handbook. The information presented here is based on the recent literature and updates previously published guidelines by Irwig and colleagues (3).

Definition of the Objectives of the Review

Diagnostic test accuracy refers to the ability of a test to distinguish between patients with disease (or more generally, a specified target condition) and those without. In a study of test accuracy, the results of the test under evaluation, the index test, are compared with those of the reference standard determined in the same patients. The reference standard is an agreed-on and accurate method for identifying patients who have the target condition. Test results are typically categorized as positive or negative for the target condition. By using such binary test outcomes, the accuracy is most often expressed as the test's sensitivity (the proportion of patients with positive results on the reference standard that are also positive on the index test) and specificity (the proportion of patients with negative results on the reference standard that are also negative on the index test). Other measures have been proposed and are in use (4–6).

It has long been recognized that test accuracy is not a fixed property of a test. It can vary between patient subgroups, with their spectrum of disease, with the clinical setting, or with the test interpreters and may depend on the results of previous testing. For this reason, inclusion of these elements in the study question is essential. In order to make a policy decision to promote use of a new index test, evidence is required that using the new test increases test accuracy over other testing options, including current practice, or that the new test has equivalent accuracy but offers other advantages (7–9). As with the evaluation of interventions, systematic reviews need to include comparative analyses between alternative testing strategies and should not

See also:

Print
Editorial comment. 904

Web-Only
Appendix
Appendix Table
Conversion of graphics into slides

focus solely on evaluating the performance of a test in isolation.

In relation to the existing situation, 3 possible roles for a new test can be defined: replacement, triage, and add-on (7). If a new test is to replace an existing test, then comparing the accuracy of both tests on the same population and with the same reference standard provides the most direct evidence. In triage, the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue the testing pathway. When a test is needed to rule out disease in patients who then need no further testing, a test that gives a minimal proportion of false–negative results and thus a relatively high sensitivity should be used. Triage tests may be less accurate than existing ones, but they have other advantages, such as simplicity or low cost. A third possible role of a new test is add-on. The new test is then positioned after the existing testing pathway to identify false-positive or false-negative results after the existing pathway. The review should provide data to assess the incremental change in accuracy made by adding the new test.

An example of a replacement question can be found in a systematic review of the diagnostic accuracy of urinary markers for primary bladder cancer (10). Clinicians may use cytology to triage patients before they undergo invasive cystoscopy, the reference standard for bladder cancer. Because cytology combines high specificity with low sensitivity (11), the goal of the review was to identify a tumor marker with sufficient accuracy to either replace cytology or be used in addition to cytology. For a marker to replace cytology, it has to achieve equally high specificity with improved sensitivity. New markers that are sensitive but not specific may have roles as adjuncts to conventional testing. The review included studies in which the test under evaluation (several different tumor markers and cytology) was evaluated against cystoscopy or histopathology. Included studies compared 1 or more of the markers, cytology only, or a combination of markers and cytology.

Although information on accuracy can help clinicians make decisions about tests, good diagnostic accuracy is a desirable but not sufficient condition for the effectiveness of a test (8). To demonstrate that using a new test does more good than harm to patients tested, randomized trials of test-and-treatment strategies and reviews of such trials may be necessary. However, with the possible exception of screening, in most cases, such randomized trials are not available and systematic reviews of test accuracy may provide the most useful evidence available to guide clinical and health policy decision making and use as input for decision and cost-effectiveness analysis (12).

Identification and Selection of Studies

Identifying test accuracy studies is more difficult than searching for randomized trials (13). There is not a clear, unequivocal keyword or indexing term for an accuracy study in literature databases comparable with the term "randomized, controlled trial." The Medical Subject Heading "sensitivity and specificity" may look suitable but is inconsistently applied in most electronic bibliographic databases. Furthermore, data on diagnostic test accuracy may be hidden in studies that did not have test accuracy estimation as their primary objective. This complicates the efficient identification of diagnostic test accuracy studies in electronic databases, such as MEDLINE. Until indexing systems properly code studies of test accuracy, searching for them will remain challenging and may require additional manual searches, such as screening reference lists.

In the development of a comprehensive search strategy, review authors can use search strings that refer to the test(s) under evaluation, the target condition, and the patient description or a subset of these. For tests with a clear name that are used for a single purpose, searching for publications in which those tests are mentioned may suffice. For other reviews, adding the patient description may be necessary, although this is also often poorly indexed. A search strategy in MEDLINE should contain both Medical Subject Headings and free text words. A search strategy for articles about tests for bladder cancer, for example, should include as many synonyms for bladder cancer as possible in the search strategy, including neoplasm, carcinoma, transitional cell, and hematuria.

Several methodological electronic search filters for diagnostic test accuracy studies have been developed, each attempting to restrict the search to articles that are most likely to be test accuracy studies (13–16). These filters rely on indexing terms for research methodology and text words used in reporting results, but they often miss relevant studies and are unlikely to decrease the number of articles one needs to screen. Therefore, they are not recommended for systematic reviews (17, 18). The incremental value of searching in languages other than English and in the gray literature has not yet been fully investigated.

In systematic reviews of intervention studies, publication bias is an important and well-studied form of bias in which the decision to report and publish studies is linked to their findings. For clinical trials, the magnitude and determinants of publication bias have been identified by tracing the publication history of cohorts of trials reviewed by ethics committees and research boards (19). A consistent observation has been that studies with significant results are more likely to be published than studies with nonsignificant findings (19). Investigating publication bias for diagnostic tests is problematic, because many studies are done without ethical review or study registration; therefore, identification of cohorts of studies from registration to final publication status is not possible (20). Funnel plot–based tests used to detect publication bias in reviews of randomized, controlled trials have proven to be seriously misleading for diagnostic studies, and alternatives have poor power (21). Also, because diagnostic accuracy studies frequently do not compare tests, they tend not to routinely

report *P* values that dichotomize comparisons as significant or not significant. Without the same emphasis being given to statistical significance, the determinants for publication of diagnostic studies are unlikely to be the same as those of intervention studies.

ASSESSMENT OF METHODOLOGICAL QUALITY

Variability among diagnostic accuracy study results is to be expected. Some of this variability is due to chance, because many diagnostic studies have small sample sizes (22). The remaining heterogeneity may be due to differences in study populations, but differences in study methods are also likely to result in differences in accuracy estimates (23). Test accuracy studies with design deficiencies can produce biased results (24–26). The **Table** lists some of the more important forms of bias. Sources of bias for which unambiguous evidence indicates that they lead to overestimation of diagnostic accuracy are the inclusion of healthy control participants and the differential use of reference standards (24, 26).

Quality assessment of individual studies in systematic reviews is therefore necessary to identify potential sources of bias and to limit the effects of these biases on the estimates and the conclusions of the review. We recommend the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) checklist to assess the quality of diagnostic test accuracy studies (27). In addition, specific sources of bias may exist for different types of diagnostic tests. For example, in studies assessing the accuracy of biochemical serum markers, data-driven selection of the cutoff value may bias diagnostic accuracy (28, 29). Review authors should therefore think carefully about whether specific items need to be added to the QUADAS list.

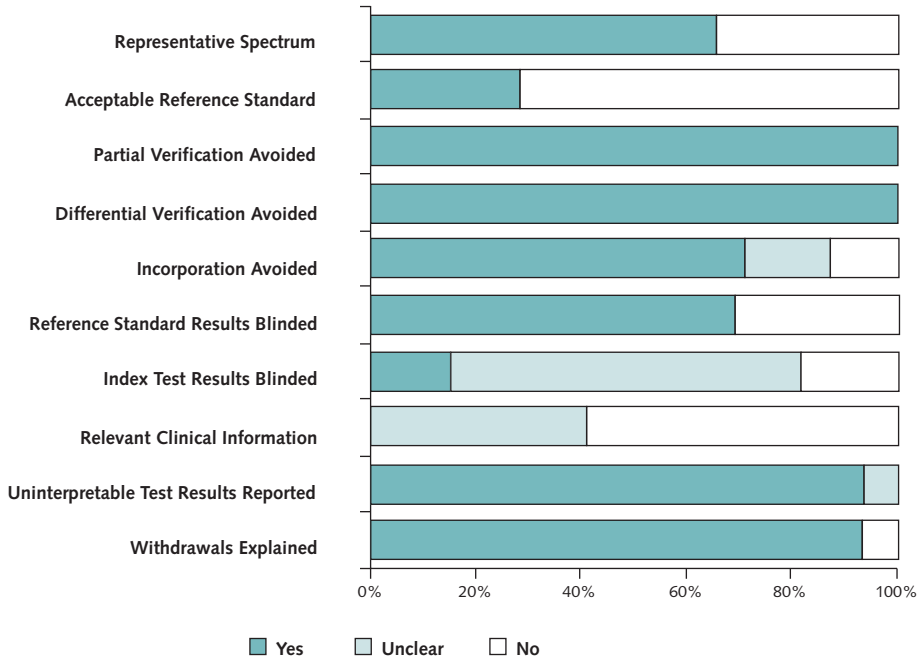
The results of quality appraisal can be summarized to offer a general impression of the validity of the available evidence. Review authors should not use an overall quality score, because different shortcomings may generate different magnitudes of bias, even in opposing directions, which makes it very hard to attach sensible weights to each quality item (30). **Figure 1** shows a way to summarize the quality assessment, with stacked bars used for each QUADAS item. Another way of presenting the quality assessment results is by tabulating the results of the individual QUADAS items for each study. In the analysis phase, the results of the quality appraisal may guide explorations of the sources of heterogeneity (32, 33). Possible methods to address quality differences are sensitivity analysis, subgroup analysis, or meta-regression analysis, although the number of included studies may often be too small for meaningful investigations. Also, incomplete reporting hampers any evaluation of study quality (34). The

Table. **Sources of Bias in Diagnostic Test Accuracy Studies**

Type of Bias	When Does It Occur?	Under- or Overestimation of Diagnostic Accuracy?*
Patients		
Spectrum bias	When included patients do not represent the intended spectrum of severity for the target condition or alternative conditions	Depends on difference between targeted and included part of spectrum
Selection bias	When eligible patients are not enrolled consecutively or randomly	Usually leads to overestimation
Index test		
Information bias	When the index test results are interpreted with knowledge of the results of the reference standard, or with more (or less) information than in practice	Usually leads to overestimation, unless less clinical information is provided than in practice, which may result in underestimation
Reference standard		
Misclassification bias	When the reference standard does not correctly classify patients with the target condition	Depends on whether both tests make the same mistakes
Partial verification bias	When a nonrandom set of patients does not undergo the reference standard	Usually leads to overestimation of sensitivity; effect on specificity varies
Differential verification bias	When a set of patients is verified with a second or third reference standard, especially when this selection depends on the index test result	Usually leads to overestimation
Incorporation bias	When the index test is incorporated in a (composite) reference standard	Usually leads to overestimation
Disease progression bias	When the patients' condition changes between administering the index test and the reference standard	Under- or overestimation, depending on change in patients' condition
Information bias	When the reference standard is interpreted knowing the index test results	Usually leads to overestimation
Data analysis		
Excluded data	When uninterpretable or intermediate test results and withdrawals are not included in the analysis	Usually leads to overestimation

* From references 24–26.

Figure 1. **Review authors' judgments about quality items in a systematic review of magnetic resonance imaging for multiple sclerosis.**



Data from reference 31. Data are presented as the proportion of included studies. Criteria that are unclear or not met introduce a risk for bias. The authors considered the relative lack of an acceptable reference standard as the main weakness of the review.

effects of the STARD guidelines for complete and transparent reporting (1) are only gradually becoming visible in the literature (35).

Analyzing the Data and Presenting the Results

Whereas the results of a randomized trial are often reported by using a single measure of effect, such as a difference in means, a risk difference, or a risk ratio, most diagnostic test accuracy studies report 2 or more statistics: the sensitivity and the specificity, the positive and negative predictive value, the likelihood ratios for the respective test results, or the receiver-operating characteristic (ROC) curve and quantities based on it (6, 36).

The first step in the meta-analysis of diagnostic test accuracy is to graph the results of the individual studies. The paired results for sensitivity and specificity in the included studies should be plotted as points in ROC space (**Figure 2**), which can highlight the covariation between sensitivity and specificity. In **Figure 2**, the x-axis of the ROC plot displays the specificity obtained in the studies in the review. The y-axis shows the corresponding sensitivity. The rising diagonal line indicates values of sensitivity and specificity that could be obtained by guessing and refers to a noninformative test: The chances of a positive test result are identical for patients with disease and those without. It is expected that most studies will be above this line. The best diagnostic tests will be positioned in the upper-left corner of the ROC space, where both sensitivity and specificity are close to 1. Because CIs are not typically displayed on these plots, it is not possible to discern the cause of scatter across studies—it can be caused by either small sample sizes or heterogeneity between studies. Paired forest plots (**Figure 3**) display sensitivity and specificity separately (but on the same row) for each study together with CIs and tabular data. A disadvantage is that forest plots do not display the covariation between sensitivity and specificity.

The estimated sensitivity and specificity of a test often display a pattern of negative correlation. A major contributor to this appearance is the tradeoff between sensitivity and specificity when the threshold for defining test positivity varies. When high test results are positive, decreasing the threshold value that defines a test result as positive increases sensitivity and lowers specificity, and vice versa. When studies included in a review differ in positivity thresholds, an ROC-curve–like pattern may be discerned in the ROC plot. There may be explicit variation in thresholds if different studies use different numerical thresholds to define a test result as positive (for example, variation in the blood glucose level, above which a patient has diabetes). In other situations, unquantifiable or implicit variation in threshold may occur when test results depend on interpretation or judgment (for example, between radiographers classifying images as normal or abnormal) or when test results are sensitive to machine calibration.

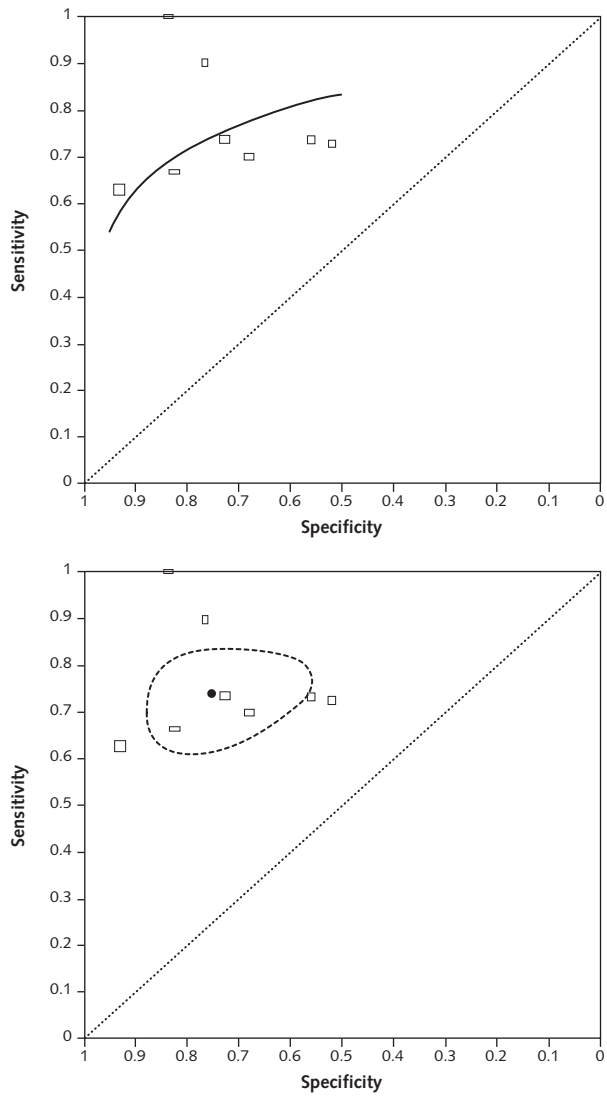
Because threshold effects cause sensitivity and specificity estimates to seem negatively correlated, and because threshold variation can be expected in many situations, robust approaches to meta-analysis take the underlying relationship between sensitivity and specificity into account. One way of doing so is by constructing a summary ROC curve. An average sensitivity and specificity point on this curve indicates where the center of the study results is. Separate pooling of sensitivity and specificity to identify this point has been discredited, because such an approach may identify a summary point that is not representative of the paired data (for example, a point that does not lie on the summary ROC curve).

Meta-analyses of studies reporting pairs of sensitivity and specificity estimates have often used the linear regression model for the construction of summary ROC curves proposed by Moses and colleagues (51), which is based on regressing the log diagnostic odds ratio against a measure of the proportion reported as positive. To examine differences between tests and to relate them to study or sample characteristics, the regression model can be extended by adding covariates (52). However, we now know that the formulation of the Moses model has its limitations. It fails to consider the precision of the study estimates, does not estimate between-study heterogeneity, and the explanatory variable in the regression is measured with error. These problems render estimates of CIs and *P* values unsuitable for formal inference (36, 53).

Two newly developed approaches to fitting random effects in hierarchical models overcome these limitations: the hierarchical summary ROC model (36, 54–56) and the bivariate random-effects model (53, 57). Both approaches model the distribution of the observed pairs of sensitivity and specificity values from each study. The hierarchical summary ROC model assumes an explicit formula linking sensitivity and specificity through a threshold; accounts for the variability across studies; and can be used to estimate summaries of the data, including a summary ROC curve and average values of accuracy measures, such as sensitivity and specificity. The bivariate random-effects model focuses on estimating the average sensitivity and specificity, but also estimates the unexplained variation in these parameters and the correlation between them. These 2 basic models are mathematically equivalent in the absence of covariates (58). Both models give a valid estimation of the underlying summary ROC curve and the average sensitivity and specificity (53, 58). Addition of covariates to the models, or application of separate models to different subgroups, enables exploration of heterogeneity. Both models can be fitted with statistical software for fitting mixed models (36, 53, 55, 57).

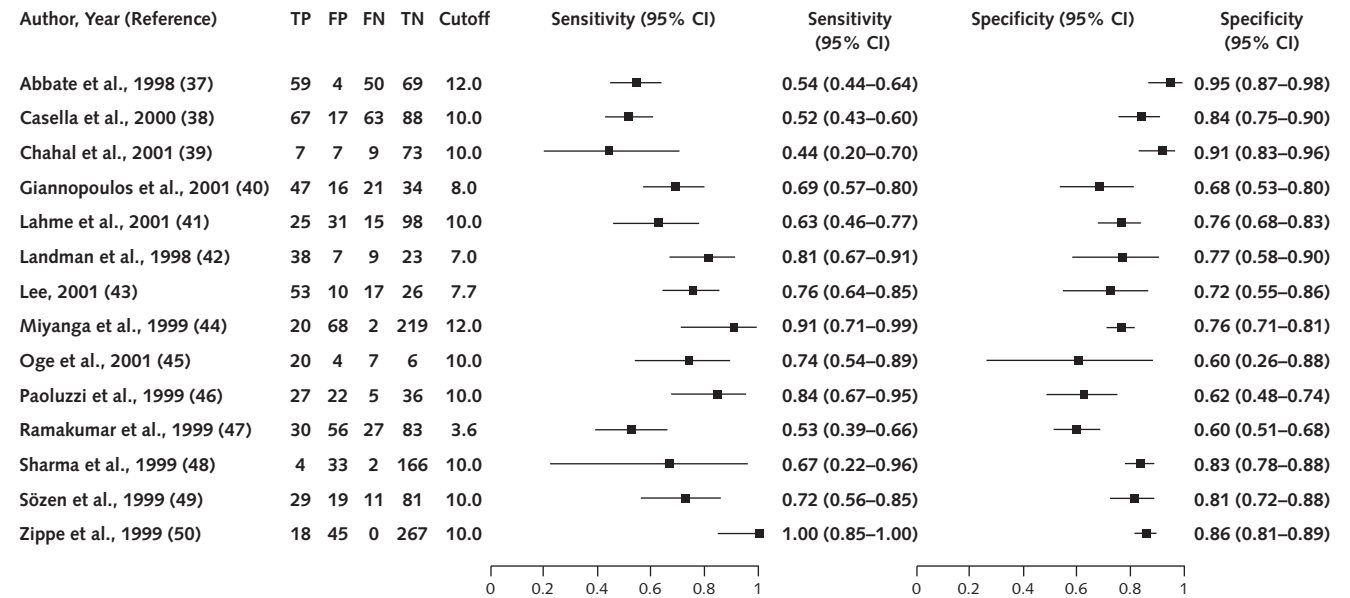
Estimates of summary likelihood ratios can best be derived from summary estimates of sensitivity and specificity obtained by using the methods described previously. Although some authors have advocated pooling likelihood ratios rather than sensitivity and specificity or ROC curves

Figure 2. **Summary receiver-operating characteristic (*ROC*) curve plots showing test accuracy of a tumor marker for bladder cancer from 8 studies included in a systematic review.**



Data from reference 10. Each study is represented by a small box positioned at the estimated sensitivity and specificity. The height and width of each box are proportional to the numbers of patients with and without bladder cancer, respectively, in each study. **Top.** This panel shows the summary ROC curve that can be drawn through these values. The scatter of the points fit, to a degree, with the existence of a threshold-type relationship between sensitivity and specificity. The curve is an estimate of the underlying relationship between sensitivity and specificity for the test used across varying thresholds. **Bottom.** This panel shows the average sensitivity and specificity estimate of the study results (*solid circle*) and a 95% confidence region around it. Estimation of a summary point only makes sense when the included studies have used a common threshold. The curves, points, and confidence regions can be estimated by using either the hierarchical summary ROC curve model (36, 54–56) or the bivariate random-effects model (53, 57).

Figure 3. **Paired forest plot of the sensitivity and specificity of a tumor marker for bladder cancer.**



FN = false-negative; FP = false-positive; TN = true-negative; TP = true-positive. Data are from reference 10. Forest plots document the extracted data for each study (numbers of TP, FP, FN, and TN results) together with estimates of sensitivity and specificity accompanied by 95% CIs. The scatter of the estimates and CIs indicates that the variability in sensitivity and specificity is unlikely to be explained by chance only, but it is not possible to ascertain whether a threshold-type relationship is evident.

(59–61), these methods do not account for the correlated bivariate nature of likelihood ratios and may yield impossible summary estimates and CIs, with positive and negative likelihood ratios either both above or both below 1.0 (62).

ROC Curves and Summary Estimates

The ability to estimate underlying summary ROC curves and average sensitivities and specificities allows flexibility in testing hypotheses and estimating diagnostic accuracy. Analyses based on all included studies facilitate well-powered comparisons between different tests or between subgroups of studies, which are not restricted to investigating accuracy at a particular threshold. The top panel of **Figure 2** shows such a summary ROC curve for the diagnostic accuracy of a tumor antigen test for diagnosing bladder cancer. In contrast, when a test is being used at the same threshold in all included studies, review authors may make a summary estimate of sensitivity and specificity. The uncertainty associated with the estimate can be described by confidence regions marked on the summary ROC plot around the average point. The bottom panel of **Figure 2** illustrates this approach.

Judgments about the validity of pooling data should be informed by considering the quality of the studies, the similarity of patients and tests being pooled, and whether the results may consequently be misleading. Where there is statistical heterogeneity in results, random-effects models are used to account for the variability and to derive suitably conservative assessments of the uncertainty in the estimates. Naturally, increased uncertainty about the estimates may make it more difficult to draw firm conclusions about the accuracy of a particular test.

Comparative Analyses

Systematic reviews of diagnostic test accuracy may evaluate more than 1 test to determine which test or combination of tests can better serve the intended purpose. Indirect comparisons can be made by calculating separate summary estimates of the sensitivity and specificity for each test, including all studies that have evaluated that test regardless of whether they evaluated the other tests. The substantial variability that can be expected between tests means that such comparisons are prone to confounding. Restricting inclusion to studies of similar design and patient characteristics may limit confounding. A theoretically preferable approach is to use only studies that have directly compared the tests in the same patients or have randomly assigned patients to 1 of the tests. Such direct comparisons do not suffer from confounding. Paired analyses can be displayed in a ROC plot, by linking the sensitivity–specificity pairs from each study with a dashed line, as in **Figure 4**. Unfortunately, fully paired studies are not always available.

Interpretation of the Results

The interpretation of the results offered in the systematic review should help readers to understand the implications for practice. This interpretation should consider whether evidence derived from the review suitably ad-

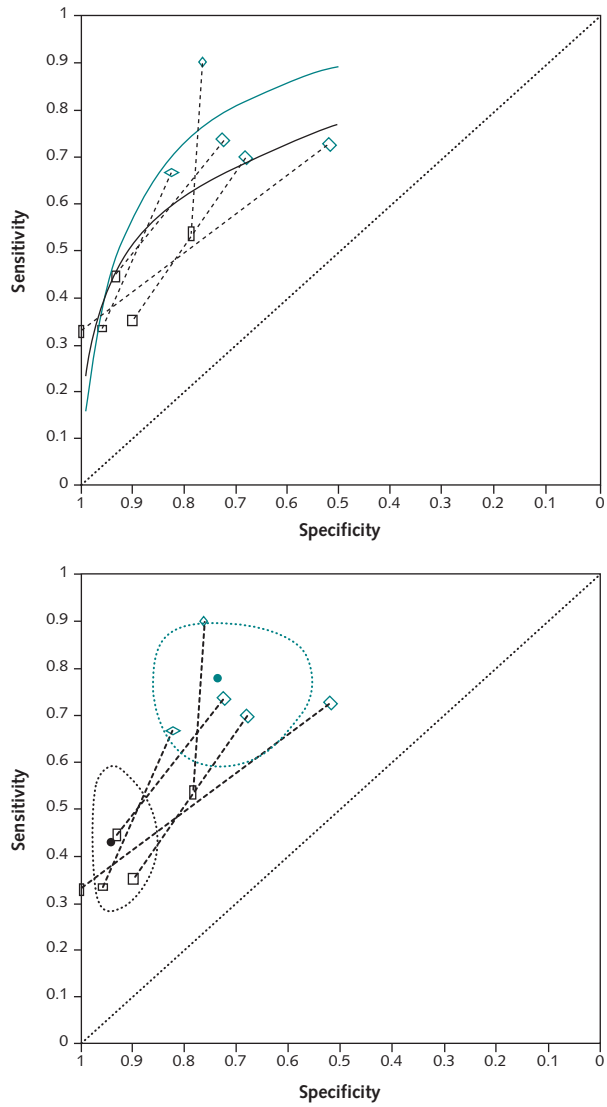
dresses the objectives of the review. It may involve considerations about whether the study sample was representative, the included studies indeed investigated the intended future role of the test under evaluation, and the results are unlikely to be biased. Review authors should consider the potential effects of quality differences on the results or the lack of high-quality studies. The interpretation of the findings should also consider the consequences of the false-positive and false-negative results and whether the estimates of accuracy are sufficiently high for the foreseen role that the test will have in practice. Some reviews may not result in useful summary estimates of sensitivity and specificity, for example, because of large variability in the individual study estimates. A decision model could be used to structure the interpretation of the findings. Such a model would incorporate important factors, such as the disease prevalence, probable outcomes, and available diagnostic and therapeutic interventions that may follow the test. Additional information, such as costs or important tradeoffs between harms and benefits, can be included (12).

CONCLUSION

The development of the methodology for systematic reviews of diagnostic test accuracy studies has made important progress in recent years. We now know more about searching, sources of bias in study design, quality appraisal, and data analysis. In meta-analysis, new hierarchical random-effects models have been developed with sound statistical properties that allow robust inferences. Methods for the estimation of summary ROC curves and summary estimates of sensitivity and specificity are now available. All these advances will be described in detail in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (63). The **Appendix Table** (available at www.annals.org) provides a summary of the key issues that both readers and review authors should consider.

Diagnostic test accuracy reviews face 2 major challenges. First, they are limited by the quality and availability of primary test accuracy studies that address important relevant questions. More studies are needed that recruit suitable spectrums of participants, make direct comparisons between tests, use rigorous methodology, and clearly report their methods and findings. Second, more development is needed in the area of interpretation and presentation of the results of diagnostic test accuracy reviews. Clinicians struggle with the definitions of sensitivity, specificity, and likelihood ratios (64, 65) possibly because, in the clinical context, the predictive value of tests is more immediately relevant. The results of systematic reviews of diagnostic accuracy can, of course, be used to assess the predictive value. Policymakers and guideline developers may be particularly interested in comparative accuracy, the costs and burden of testing, or new test methods. Developing systematic reviews that are relevant for policymakers and clin-

Figure 4. **Meta-analysis of the diagnostic test accuracy of 2 index tests for bladder cancer: cytology (*black squares*) and bladder tumor antigen (*green diamonds*).**



Data are from reference 10. The meta-analysis is restricted to studies that made a direct comparison between the tests by using both tests in each patient and comparing them with the reference standard of invasive cystoscopy. Restriction of the meta-analysis to direct test comparisons reduces concerns of confounding and allows stronger inferences to be drawn from the comparison of tests. The dashed lines link together the cytology and bladder tumor antigen results from each study and give the impression that bladder tumor antigen is much more sensitive but less specific than cytology. **Top.** Summary receiver-operating characteristic curves fitted to the data indicate that the bladder tumor antigen curve dominates the cytology curve as specificity decreases. Thus, bladder tumor antigen has the potential to be a more sensitive test than cytology, but only at specificities below 90%. **Bottom.** Cytology has an average sensitivity of 0.43 and an average specificity of 0.94 (*black circle*); bladder tumor antigen has an average sensitivity of 0.78 and an average specificity of 0.74 (*green circle*). The nonoverlapping 95% confidence regions indicate that the differences between the tests are unlikely to have occurred by chance alone.

ical practice poses a major challenge and requires clear thinking about the scope and purpose of the review.

From the Dutch Cochrane Centre and Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; Unit of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, United Kingdom; and Center for Statistical Sciences, Brown University, Providence, Rhode Island.

Grant Support: By the UK National Institute for Health Research (grant RNC/018/0003), the National Cancer Institute (grant 2U01CA079778), and the Cochrane Collaboration.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Jonathan J. Deeks, PhD, Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom; e-mail, j.deeks@bham.ac.uk.

Current author addresses are available at www.annals.org.

References

1. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for Reporting of Diagnostic Accuracy.** Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138:40-4. [PMID: 12513043]
2. **The Cochrane Collaboration.** The Cochrane Manual Issue 3, 2008 [updated 15 May 2008]. Oxford, UK: Cochrane Collaboration; 2008. Accessed at www.cochrane.org/admin/manual.htm on 18 July 2008.
3. **Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al.** Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120:667-76. [PMID: 8135452]
4. **Knottnerus JA, ed.** The Evidence Base of Clinical Diagnosis. London: BMJ Books; 2002.
5. **Bossuyt PM.** Interpreting diagnostic test accuracy studies. Semin Hematol. 2008;45:189-95. [PMID: 18582626]
6. **Zhou X-H, Obuchowski N, McClish D.** Statistical Methods in Diagnostic Medicine. Hoboken, NJ: J Wiley; 2002.
7. **Bossuyt PM, Irwig L, Craig J, Glasziou P.** Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089-92. [PMID: 16675820]
8. **Lord SJ, Irwig L, Simes RJ.** When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850-5. [PMID: 16754927]
9. **Thornbury JR.** Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol. 1994;162:1-8. [PMID: 8273645]
10. **Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH.** Tumor markers in the diagnosis of primary bladder cancer. A systematic review. J Urol. 2003;169:1975-82. [PMID: 12771702]
11. **Lokeshwar VB, Selzer MG.** Urinary bladder tumor markers. Urol Oncol. 2006;24:528-37. [PMID: 17138134]
12. **Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein M.** Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge, UK: Cambridge Univ Pr; 2001.
13. **Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC.** Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994;1:447-58. [PMID: 7850570]
14. **Devillé WL, Bezemer PD, Bouter LM.** Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol. 2000;53:65-9. [PMID: 10693905]
15. **Bachmann LM, Coray R, Estermann P, Ter Riet G.** Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc. 2002;9:653-8. [PMID: 12386115]
16. **Haynes RB, Wilczynski NL.** Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004;328:1040. [PMID: 15073027]
17. **Doust JA, Pietrzak E, Sanders S, Glasziou PP.** Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol. 2005;58:444-9. [PMID: 15845330]
18. **Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM.** Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol. 2006;59:234-40. [PMID: 16488353]
19. **Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ.** Publication and related biases. Health Technol Assess. 2000;4:1-115. [PMID: 10932019]
20. **Song F, Khan KS, Dinnes J, Sutton AJ.** Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol. 2002;31:88-95. [PMID: 11914301]
21. **Deeks JJ, Macaskill P, Irwig L.** The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58:882-93. [PMID: 16085191]
22. **Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM.** Sample sizes of studies on diagnostic accuracy: literature survey. BMJ. 2006;332:1127-9. [PMID: 16627488]
23. **Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J.** Designing studies to ensure that estimates of test accuracy are transferable. BMJ. 2002;324:669-71. [PMID: 11895830]
24. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]
25. **Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140:189-202. [PMID: 14757617]
26. **Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM.** Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76. [PMID: 16477057]
27. **Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J.** The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25. [PMID: 14606960]
28. **Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH.** Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem. 2008;54:729-37. [PMID: 18258670]
29. **Ewald B.** Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006;59:798-801. [PMID: 16828672]
30. **Whiting P, Harbord R, Kleijnen J.** No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol. 2005;5:19. [PMID: 15918898]
31. **Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, et al.** Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. BMJ. 2006;332:875-84. [PMID: 16565096]
32. **Westwood ME, Whiting PF, Kleijnen J.** How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol. 2005;5:20. [PMID: 15943861]
33. **Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, et al.** Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. Clin Chem. 2007;53:164-72. [PMID: 17185365]
34. **Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al.** Quality of reporting of diagnostic accuracy studies. Radiology. 2005;235:347-53. [PMID: 15770041]
35. **Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al.** The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology. 2006;67:792-7. [PMID: 16966539]
36. **Gatsonis C, Paliwal P.** Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol. 2006;187:271-81. [PMID: 16861527]
37. **Abbate I, D'Introno A, Cardo G, Marano A, Addabbo L, Musci MD, et al.** Comparison of nuclear matrix protein 22 and bladder tumor antigen in urine of patients with bladder cancer. Anticancer Res. 1998;18:3803-5. [PMID: 9854500]
38. **Casella R, Huber P, Blöchlinger A, Stoffel F, Dalquen P, Gasser TC, et al.** Urinary level of nuclear matrix protein 22 in the diagnosis of bladder cancer:

experience with 130 patients with biopsy confirmed tumor. J Urol. 2000;164: 1926-8. [PMID: 11061883]
39. **Chahal R, Darshane A, Browning AJ, Sundaram SK.** Evaluation of the clinical value of urinary NMP22 as a marker in the screening and surveillance of transitional cell carcinoma of the urinary bladder. Eur Urol. 2001;40:415-20; discussion 421. [PMID: 11713396]
40. **Giannopoulos A, Manousakas T, Gounari A, Constantinides C, Choremi-Papadopoulou H, Dimopoulos C.** Comparative evaluation of the diagnostic performance of the BTA stat test, NMP22 and urinary bladder cancer antigen for primary and recurrent bladder tumors. J Urol. 2001;166:470-5. [PMID: 11458049]
41. **Lahme S, Bichler KH, Feil G, Krause S.** Comparison of cytology and nuclear matrix protein 22 for the detection and follow-up of bladder cancer. Urol Int. 2001;66:72-7. [PMID: 11223747]
42. **Landman J, Chang Y, Kavaler E, Droller MJ, Liu BC.** Sensitivity and specificity of NMP-22, telomerase, and BTA in the detection of human bladder cancer. Urology. 1998;52:398-402. [PMID: 9730450]
43. **Lee KH.** Evaluation of the NMP22 test and comparison with voided urine cytology in the detection of bladder cancer. Yonsei Med J. 2001;42:14-8. [PMID: 11293494]
44. **Miyanaga N, Akaza H, Tsukamoto T, Ishikawa S, Noguchi R, Ohtani M, et al.** Urinary nuclear matrix protein 22 as a new marker for the screening of urothelial cancer in patients with microscopic hematuria. Int J Urol. 1999;6: 173-7. [PMID: 10226833]
45. **Oge O, Atsü N, Kendi S, Ozen H.** Evaluation of nuclear matrix protein 22 (NMP22) as a tumor marker in the detection of bladder cancer. Int Urol Nephrol. 2001;32:367-70. [PMID: 11583354]
46. **Paoluzzi M, Cuttano MG, Mugnaini P, Salsano F, Giannotti P.** Urinary dosage of nuclear matrix protein 22 (NMP22) like biologic marker of transitional cell carcinoma (TCC): a study on patients with hematuria. Arch Ital Urol Androl. 1999;71:13-8. [PMID: 10193018]
47. **Ramakumar S, Bhuiyan J, Besse JA, Roberts SG, Wollan PC, Blute ML, et al.** Comparison of screening methods in the detection of bladder cancer. J Urol. 1999;161:388-94. [PMID: 9915409]
48. **Sharma S, Zippe CD, Pandrangi L, Nelson D, Agarwal A.** Exclusion criteria enhance the specificity and positive predictive value of NMP22 and BTA stat. J Urol. 1999;162:53-7. [PMID: 10379739]
49. **Sözen S, Biri H, Sinik Z, Küpeli B, Alkibay T, Bozkirli I.** Comparison of the nuclear matrix protein 22 with voided urine cytology and BTA stat test in the diagnosis of transitional cell carcinoma of the bladder. Eur Urol. 1999;36:225-9. [PMID: 10450007]
50. **Zippe C, Pandrangi L, Agarwal A.** NMP22 is a sensitive, cost-effective test in patients at risk for bladder cancer. J Urol. 1999;161:62-5. [PMID: 10037369]
51. **Moses LE, Shapiro D, Littenberg B.** Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12:1293-316. [PMID: 8210827]
52. **Lijmer JG, Bossuyt PM, Heisterkamp SH.** Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med. 2002;21:1525-37. [PMID: 12111918]
53. **Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T.** Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28:621-638. [PMID: 18591542]
54. **Rutter CM, Gatsonis CA.** A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84. [PMID: 11568945]
55. **Macaskill P.** Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol. 2004;57:925-32. [PMID: 15504635]
56. **Dukic V, Gatsonis C.** Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003;59:936-46. [PMID: 14969472]
57. **Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH.** Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982-90. [PMID: 16168343]
58. **Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA.** A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239-51. [PMID: 16698768]
59. **Stengel D, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F.** A likelihood ratio approach to meta-analysis of diagnostic studies. J Med Screen. 2003;10:47-51. [PMID: 12790315]
60. **Khan KS.** Systematic reviews of diagnostic tests: a guide to methods and application. Best Pract Res Clin Obstet Gynaecol. 2005;19:37-46. [PMID: 15749064]
61. **Khan KS, Dinnes J, Kleijnen J.** Systematic reviews to evaluate diagnostic tests. Eur J Obstet Gynecol Reprod Biol. 2001;95:6-11. [PMID: 11267714]
62. **Zwinderman AH, Bossuyt PM.** We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med. 2008;27:687-97. [PMID: 17611957]
63. **Deeks JJ, Bossuyt PM, Gatsonis C (editors).** Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. Oxford, UK: The Cochrane Collaboration, 2009. [Forthcoming].
64. **Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G.** Communicating accuracy of tests to general practitioners: a controlled study. BMJ. 2002;324: 824-6. [PMID: 11934776]
65. **Puhan MA, Steurer J, Bachmann LM, ter Riet G.** A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. Ann Intern Med. 2005;143:184-9. [PMID: 16061916]

Current Author Addresses: Drs. Leeflang and Bossuyt: Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, the Netherlands.

Dr. Deeks: Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.

Dr. Gatsonis: Center for Statistical Sciences, Brown University, Box G-S121, 121 South Main Street, 7th Floor, Providence, RI 02912.

Appendix: Contributors to the Cochrane Diagnostic Test Accuracy Working Group

Contributors are listed in alphabetical order.

Bert Aertgeerts, Doug Altman, Gerd Antes, Lucas Bachmann, Patrick Bossuyt, Heiner Buchner, Peter Bunting, Frank Buntinx, Jonathan Craig, Roberto D'Amico, Riekie de Vet, Jon Deeks, Jenny Doust, Matthias Egger, Anne Eisinga, Graziella Fillipini, Yngve Flack-Ytter, Constantine Gatsonis, Afina Glas, Paul Glasziou, Fritz Grossenbacher, Roger Harbord, Jorgen Hilden, Lotty Hooft, Andrea Horvath, Chris Hyde, Les Irwig, Monica Kjeldstrøm, Petra Macaskill, Susan Mallett, Ruth Mitchell, Tess Moore, Rasmus Moustgaard, Wytze Oosterhuis, Madhukar Pai, Prashni Paliwal, Daniel Pewsner, Hans Reitsma, Jacob Riis, Ingrid Riphagen, Anne Rutjes, Rob Scholten, Nynke Smidt, Jonathan Sterne, Yemisi Takwoingi, Danielle van der Windt, Vasivy Vlassov, Joseph Watine, and Penny Whiting.

Appendix Table. **Essential Elements in a Systematic Review of Diagnostic Test Accuracy**

Phase in Review Process	Key Issues
1. Definition of the review objectives	To identify the review question: State the patient group and define presenting condition(s), previous test results, and health care setting. Describe the tests (or test strategies) under evaluation, specifying their intended roles. Identify tests and test strategies currently used in practice for comparison, if available. Define the target condition to be diagnosed and reference standards to be used.
2. Study identification and selection	Search several electronic databases. Use a search strategy built around terms for the index test, target condition, and possibly patient characteristics. Do not use restrictive methodological search filters.
3. Quality assessment	Identify biases for which the included studies are at risk. Use the QUADAS checklist as a tool for identifying many common deficiencies. Comment on the adequacy of each aspect of study design. Do not use summary quality scores.
4. Data extraction, analysis, and presentation	Extract paired estimates of test sensitivity and specificity from each study overall and, if available, for patient subgroups. Plot studies in ROC space to identify the location, variability, and correlations. The hierarchical summary ROC and bivariate random-effects models provide a sound statistical framework for analysis, accounting for sampling variability, unexplained heterogeneity, and covariation between sensitivity and specificity. Compute average values of sensitivity and specificity when the data combined share a common threshold. Use summary ROC curves to describe test performance and to compare tests without restricting to particular thresholds. Obtain estimates of summary likelihood ratios from average values of sensitivity and specificity and not through separate pooling of likelihood ratios. Global tests for heterogeneity before data synthesis or tests for publication bias are typically not useful. Meta-analyze and present studies that compare tests by using randomized or within-patient designs separately from the results of indirect comparisons.
5. Interpretation	Consider the consequences of using the test, in terms of (changes in) the numbers of true-positive, false-positive, true-negative, and false-negative test results with the expected prevalence of the target disorder. Address the applicability of the results in terms of whether the patients in the primary studies were similar to those outlined in the objective, and whether tests and test strategies evaluated and compared were representative of test strategies that are used in practice. Address to what extent the original studies were biased and how these biases could influence the results and the degree to which comparisons between tests may be confounded. Consider complementing the interpretation with decision modeling by using results of the review.

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; ROC = receiver-operating characteristic.

Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative

Patrick M. Bossuyt, Johannes B. Reitsma, David E. Bruns, Constantine A. Gatsonis, Paul P. Glasziou, Les M. Irwig, Jeroen G. Lijmer, David Moher, Drummond Rennie, and Henrica C.W. de Vet, for the STARD Group*

Background: **To comprehend the results of diagnostic accuracy studies, readers must understand the design, conduct, analysis, and results of such studies. That goal can be achieved only through complete transparency from authors.**

Objective: **To improve the accuracy and completeness of reporting of studies of diagnostic accuracy in order to allow readers to assess the potential for bias in the study and to evaluate its generalizability.**

Methods: **The Standards for Reporting of Diagnostic Accuracy (STARD) steering committee searched the literature to identify publications on the appropriate conduct and reporting of diagnostic studies and extracted potential items into an extensive list. Researchers, editors, methodologists and statisticians, and members of professional organizations shortened this list during a 2-day consensus meeting with the goal of developing a checklist and a generic flow diagram for studies of diagnostic accuracy.**

Results: **The search for published guidelines on diagnostic research yielded 33 previously published checklists, from which we extracted a list of 75 potential items. The consensus meeting shortened the list to 25 items, using evidence on bias whenever available. A prototypical flow diagram provides information about the method of patient recruitment, the order of test execution, and the numbers of patients undergoing the test under evaluation, the reference standard, or both.**

Conclusions: **Evaluation of research depends on complete and accurate reporting. If medical journals adopt the checklist and the flow diagram, the quality of reporting of studies of diagnostic accuracy should improve to the advantage of the clinicians, researchers, reviewers, journals, and the public.**

Ann Intern Med. 2003;138:40-44. www.annals.org

For author affiliations, see end of text.

*For members of the STARD Group, see Appendix.

See related article, available only at www.annals.org.

The world of diagnostic tests is highly dynamic. New tests are developed at a fast rate and the technology of existing tests is continuously being improved. Exaggerated and biased results from poorly designed and reported diagnostic studies can trigger their premature dissemination and lead physicians into making incorrect treatment decisions. A rigorous evaluation process of diagnostic tests before introduction into clinical practice could not only reduce the number of unwanted clinical consequences related to misleading estimates of test accuracy, but also limit health care costs by preventing unnecessary testing. Studies to determine the diagnostic accuracy of a test are a vital part in this evaluation process (1–3).

In studies of diagnostic accuracy, the outcomes from one or more tests under evaluation are compared with outcomes from the reference standard, both measured in subjects who are suspected of having the condition of interest. The term *test* refers to any method for obtaining additional information on a patient's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests, and histopathology. The condition of interest or target condition can refer to a particular disease or to any other identifiable condition that may prompt clinical actions, such as further diagnostic testing, or the initiation, modification, or termination of treatment. In this framework, the *reference standard* is considered to be the best available method for establishing the presence or absence of the condition of interest. The reference standard can be a single method, or a combination of methods, to establish the presence of the target condition. It can include laboratory tests, imaging tests, and pathology, but also dedicated clinical follow-up of subjects. The term *accuracy* refers to the amount of agreement between the information from the test under evaluation, referred to as the *index test*, and the reference standard. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver-operator characteristic (ROC) curve (4–6).

There are several potential threats to the internal and external validity of a study of diagnostic accuracy. A survey of studies of diagnostic accuracy published in four major medical journals between 1978 and 1993 revealed that the methodological quality was mediocre at best (7). However, evaluations were hampered because many reports lacked information on key elements of design, conduct, and analysis of diagnostic studies (7). The absence of critical information about the design and conduct of diagnostic studies has been confirmed by authors of meta-analyses (8, 9). As in any other type of research, flaws in study design can lead to biased results. One report showed that diagnostic studies with specific design features are associated with biased, optimistic estimates of diagnostic accuracy compared to studies without such deficiencies (10).

At the 1999 Cochrane Colloquium meeting in Rome, the Cochrane Diagnostic and Screening Test Methods Working Group discussed the low methodological quality and substandard reporting of diagnostic test evaluations. The Working Group felt that the first step to correct these problems was to improve the quality of reporting of diagnostic studies. Following the successful CONSORT (Consolidated Standards of Reporting Trials) initiative (11–13),

the Working Group aimed at the development of a checklist of items that should be included in the report of a study of diagnostic accuracy.

The objective of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative is to improve the quality of reporting of studies of diagnostic accuracy. Complete and accurate reporting allows the reader to detect the potential for bias in the study (internal validity) and to assess the generalizability and applicability of the results (external validity).

Methods

The STARD steering committee (see Appendix for membership and details) started with an extensive search to identify publications on the conduct and reporting of diagnostic studies. This search included MEDLINE, EMBASE, BIOSIS, and the methodological database from the Cochrane Collaboration up to July 2000. In addition, the steering committee members examined reference lists of retrieved articles, searched personal files, and contacted other experts in the field of diagnostic research. They reviewed all relevant publications and extracted an extended list of potential checklist items.

Subsequently, the STARD steering committee convened a 2-day consensus meeting for invited experts from the following interest groups: researchers, editors, methodologists, and professional organizations. The aim of the conference was to reduce the extended list of potential items, where appropriate, and to discuss the optimal format and phrasing of the checklist. The selection of items to retain was based on evidence whenever possible.

The meeting format consisted of a mixture of small group sessions and plenary sessions. Each small group focused on a group of related items of the list. The suggestions of the small groups were then discussed in plenary sessions. Overnight, a first draft of the STARD checklist was assembled based on the suggestions from the small group and the additional remarks from the plenary sessions. All meeting attendees discussed this version the next day and made additional changes. The members of the STARD group could suggest further changes through a later round of comments by electronic mail.

Potential users field-tested the conference version of the checklist and flow diagram and additional comments were collected. This version was placed on the CONSORT Web site with a call for comments. The STARD steering committee discussed all comments and assembled the final checklist.

Results

The search for published guidelines for diagnostic research yielded 33 lists. Based on these published guidelines and on input of steering and STARD group members, the steering group assembled a list of 75 items. During the consensus meeting on 16 and 17 September 2000, participants consolidated and eliminated items to form the 25-item checklist. Conference members made major revisions to the phrasing and format of the checklist.

The STARD group received valuable comments and remarks during the various stages of evaluation after the conference, which resulted in the version of the STARD checklist that appears in the **Table**.

The flow diagram provides information about the method of patient recruitment (e.g., based on a consecutive series of patients with specific symptoms, case–control), the order of test execution, and the number of patients undergoing the test under evaluation (index test) and the reference test (**Figure**). We provide one prototypical flow chart that reflects the most commonly employed design in diagnostic research. Examples that reflect other designs are on the STARD Web site (see www.consort-statement.org \stardstatement.htm).

Discussion

The purpose of the STARD initiative is to improve the quality of the reporting of diagnostic studies. The items in the checklist and the flow chart can help authors in describing essential elements of the design and conduct of the study, the execution of tests, and the results.

We arranged the items under the usual headings of a medical research article but this is not intended to dictate the order in which they have to appear within an article.

The guiding principle in the development of the STARD checklist was to select items that would help readers to judge the potential for bias in the study and to appraise the applicability of the findings. Two other general considerations shaped the content and format of the checklist. First, the STARD group believes that one general checklist for studies of diagnostic accuracy, rather than different checklists for each field, is likely to be more widely disseminated and perhaps accepted by authors, peer reviewers, and journal editors. Although the evaluation of imaging tests differs from that of tests in the laboratory, we felt that these differences were more in degree than of kind. The second consideration was the development of a checklist specifically aimed at studies of diagnostic accuracy. We did not include general issues in the reporting of research findings, like the recommendations contained in the Uniform Requirements for Manuscripts Submitted to Biomedical Journals (14).

Wherever possible, the STARD group based the decision to include an item on evidence linking the item to biased estimates (internal validity) or to variation in measures of diagnostic accuracy (external validity). The evidence varied from narrative articles explaining theoretical principles and papers presenting results from statistical modeling to empirical evidence derived from diagnostic studies. For several items, the evidence is rather limited.

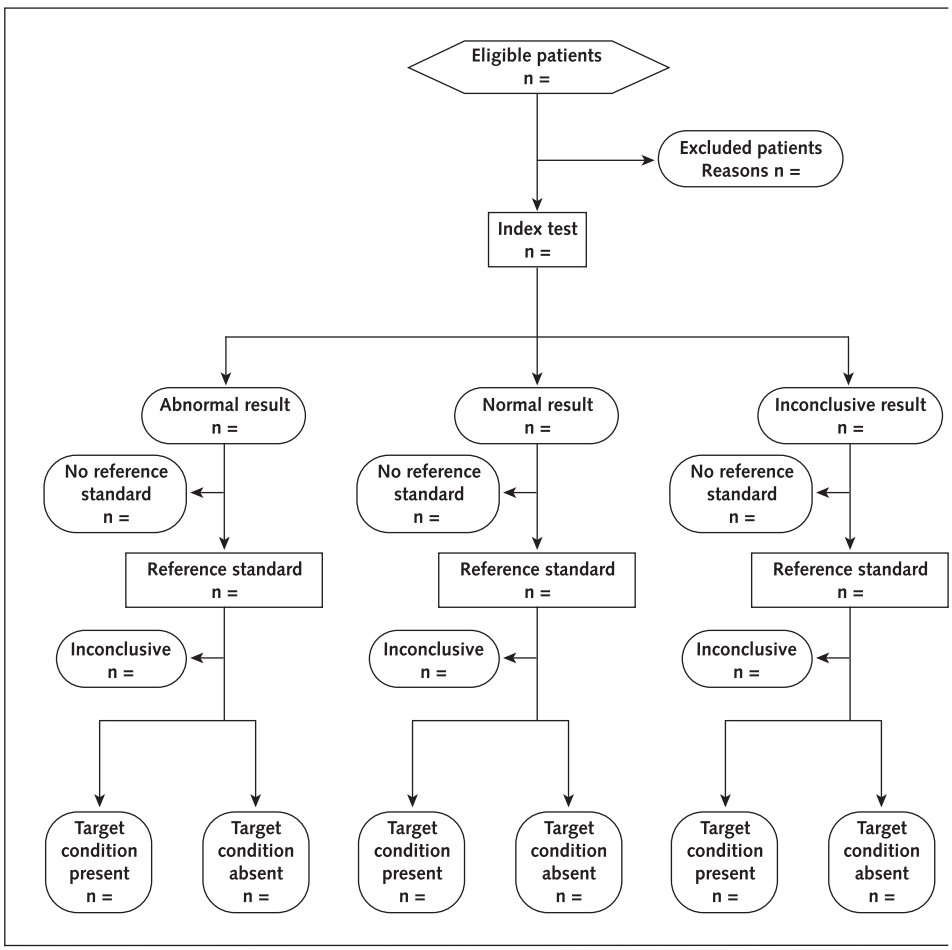
A separate background document, available at www.annals.org, explains the meaning and rationale of each item and briefly summarizes the type and amount of evidence (15). This background document should enhance

Table. **STARD Checklist for the Reporting of Studies of Diagnostic Accuracy***

Section and Topic	Item #		On page #
TITLE/ABSTRACT/KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard.	
	10	The number, training, and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).	
	22	How indeterminate results, missing responses, and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

* MeSH = Medical Subject Heading; STARD = Standards for Reporting of Diagnostic Accuracy.

Figure. **Prototypical flow diagram of a diagnostic accuracy study.**



the use, understanding, and dissemination of the STARD checklist.

The STARD group put considerable effort into the development of a flow diagram for diagnostic studies. A flow diagram has the potential to communicate vital information about the design of a study and the flow of participants in a transparent manner (16). A comparable flow diagram has become an essential element in the CONSORT standards for reporting of randomized trials (12, 16). The flow diagram could be even more essential in diagnostic studies, given the variety of designs employed in diagnostic research. Flow diagrams in the reports of diagnostic accuracy studies indicate the process of sampling and selecting participants (external validity), the flow of participants in relation to the timing and outcomes of tests, the number of subjects who fail to receive either the index test and/or the reference standard (potential for verification bias [17–19]), and the number of patients at each stage of the study, thus providing the correct denominator for proportions (internal consistency).

The STARD group plans to measure the impact of the statement on the quality of published reports on diagnostic accuracy using a before-and-after evaluation (13). Updates of STARD will be provided when new evidence on sources of bias or variability becomes available. We welcome any comments, whether on content or form, to improve the current version.

Appendix

Members of the STARD Steering Committee

Patrick Bossuyt, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands; David Bruns, *Clinical Chemistry*, Washington, D.C., United States of America; Constantine Gatsonis, Brown University, Centre for Statistical Sciences, Providence, Rhode Island, United States of America; Paul Glasziou, Mayne Medical School, Department of Social and Preventive Medicine, Herston, Australia; Les Irwig, University of Sydney, Department of Public Health and Community Medicine, Sydney, Australia; Jeroen Lijmer, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands; David Moher, Chalmers Research Group, Ottowa, Ontario, Canada; Drummond Rennie, *Journal of the American Medical Association*, Chicago, Illinois, United States of America; and Riekie de Vet, Free University, Institute for Research in Extramural Medicine, Amsterdam, the Netherlands.

Members of the STARD Group

Doug Altman, Institute of Health Sciences, Centre for Statistics in Medicine, Oxford, United Kingdom; Stuart Barton, *British Medical Journal*, BMA House, London, United Kingdom; Colin Begg,

Memorial Sloan-Kettering Cancer Center, Department of Epidemiology and Biostatistics, New York, New York, United States of America; William Black, Dartmouth-Hitchcock Medical Center, Department of Radiology, Lebanon, New Hampshire, United States of America; Harry Büller, Academic Medical Center, Department of Vascular Medicine, Amsterdam, the Netherlands; Gregory Campbell, U.S. Food and Drug Administration, Center for Devices and Radiological Health, Rockville, Maryland, United States of America; Frank Davidoff, *Annals of Internal Medicine*, Philadelphia, Pennsylvania, United States of America; Jon Deeks, Institute of Health Sciences, Centre for Statistics in Medicine, Old Road, United Kingdom; Paul Dieppe, Department of Social Medicine, University of Bristol, Bristol, United Kingdom; Kenneth Fleming, John Radcliffe Hospital, Oxford, United Kingdom; Rijk van Ginkel, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands; Afina Glas, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands; Gordon Guyatt, McMaster University, Clinical Epidemiology and Biostatistics, Hamilton, Canada; James Hanley, McGill University, Department of Epidemiology and Biostatistics, Montreal, Canada; Richard Horton, *The Lancet*, London, United Kingdom; Myriam Hunink, Erasmus Medical Center, Department of Epidemiology and Biostatistics, Rotterdam, the Netherlands; Jos Kleijnen, National Health Services Centre for Reviews and Dissemination, York, United Kingdom; Andre Knottnerus, Maastricht University, Netherlands School of Primary Care Research, Maastricht, the Netherlands; Erik Magid, Amager Hospital, Department of Clinical Biochemistry, Copenhagen, Denmark; Barbara McNeil, Harvard Medical School, Department of Health Care Policy, Boston, Massachusetts, United States of America; Matthew McQueen, Hamilton Civic Hospitals, Department of Laboratory Medicine, Hamilton, Canada; Andrew Onderdonk, Channing Laboratory, Boston, Massachusetts, United States of America; John Overbeke, Nederlands Tijdschrift voor Geneeskunde, Amsterdam, the Netherlands; Christopher Price, St. Bartholomew's—Royal London School of Medicine and Dentistry, London, United Kingdom; Anthony Proto, Radiology Editorial Office, Richmond, United States of America; Hans Reitsma, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands; David Sackett, Trout Research and Education Centre, Irish Lake, Ontario, Canada; Gerard Sanders, Academic Medical Center, Department of Clinical Chemistry, Amsterdam, the Netherlands; Harold Sox, *Annals of Internal Medicine*, Philadelphia, Pennsylvania, United States of America; Sharon Straus, Mt. Sinai Hospital, Toronto, Canada; and Stephan Walter, McMaster University, Clinical Epidemiology and Biostatistics, Hamilton, Canada.

From Academic Medical Center, University of Amsterdam, and VU University Medical Center, Amsterdam, the Netherlands; *Clinical Chemistry*, Washington, D.C.; Brown University, Providence, Rhode Island; University of Queensland Medical School, Herston, and University of Sydney, Sydney, Australia; Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada; and *Journal of the American Medical Association*, Chicago, Illinois.

Funding/Support: Financial support to convene the STARD Group was provided in part by the Dutch Health Care Insurance Board, Amstelveen, the Netherlands; the International Federation of Clinical Chemistry, Milano, Italy; the Medical Research Council's Health Services Research Collaboration, Bristol, England; and the Academic Medical Center, Amsterdam, the Netherlands.

Acknowledgment: This initiative to improve the reporting of studies of diagnostic accuracy was supported by a large number of people around the globe who commented on earlier versions.

Requests for Single Reprints: Customer Service, American College of Physicians–American Society of Internal Medicine, 190 N. Independence Mall West, Philadelphia, PA 19106.

Current author addresses are available at www.annals.org.

References

1. **Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M.** A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986;134:587-94. [PMID: 3512062]

2. **Fryback DG, Thornbury JR.** The efficacy of diagnostic imaging. Med Decis Making. 1991;11:88-94. [PMID: 1907710]

3. **Kent DL, Larson EB.** Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol. 1992;27:245-54. [PMID: 1551777]

4. **Griner PF, Mayewski RJ, Mushlin AI, Greenland P.** Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med. 1981;94:557-92. [PMID: 6452080]

5. **Sackett DL, Haynes RB, Guyatt GH, Tugwell P.** The selection of diagnostic tests. In: Sackett D, ed. Clinical Epidemiology. 2nd ed. Boston/Toronto/London: Little, Brown; 1991:47-57.

6. **Metz CE.** Basic principles of ROC analysis. Semin Nucl Med. 1978;8:283-98. [PMID: 112681]

7. **Reid MC, Lachs MS, Feinstein AR.** Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA. 1995;274:645-51. [PMID: 7637146]

8. **Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM.** Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. Radiology. 2000;217:105-14. [PMID: 11012430]

9. **de Vries SO, Hunink MG, Polak JF.** Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. Acad Radiol. 1996;3:361-9. [PMID: 8796687]

10. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]

11. **Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al.** Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276:637-9. [PMID: 8773637]

12. **Moher D, Schulz KF, Altman D.** The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285:1987-91. [PMID: 11308435]

13. **Moher D, Jones A, Lepage L.** Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA. 2001;285:1992-5. [PMID: 11308436]

14. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. JAMA. 1997;277:927-34. [PMID: 9062335] Also available at www.acponline.org.

15. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.** The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003;49:7-18.

16. **Egger M, Jüni P, Bartlett C.** Value of flow diagrams in reports of randomized controlled trials. JAMA. 2001;285:1996-9. [PMID: 11308437]

17. **Knottnerus JA.** The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making. 1987;7:139-48. [PMID: 3613914]

18. **Panzer RJ, Suchman AL, Griner PF.** Workup bias in prediction research. Med Decis Making. 1987;7:115-9. [PMID: 3574021]

19. **Begg CB.** Biases in the assessment of diagnostic tests. Stat Med. 1987;6:411-23. [PMID: 3114858]

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

In this fourth article of the five part series, we describe how guideline developers are using GRADE to rate the quality of evidence and move from evidence to a recommendation for diagnostic tests and strategies. Although recommendations on diagnosis share the fundamental logic of recommendations for other interventions, they present unique challenges. We will describe why guideline panels should be cautious when they use evidence of the accuracy of tests ("test accuracy") as the basis for recommendations and why evidence of test accuracy often provides low quality evidence for making recommendations.

AHolger J Schünemann professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Rome, Italy and CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

Andrew D Oxman researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo, Norway

Jan Brozek research fellow, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Rome, Italy

Paul Glasziou professor, Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford OX3 7LF

Roman Jaeschke clinical professor, Department of Medicine, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5

Gunn E Vist researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo, Norway

John W Williams Jr professor, Department of Medicine, Duke University and Durham VA Medical Center, Durham, NC 27705, USA

Regina Kunz associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland

Jonathan Craig associate professor, Screening and Test Evaluation Program, School of Public Health, University of Sydney, Department of Nephrology, Children's Hospital at Westmead, Sydney, Australia

Authors continued on next page

This is the fourth in a series of five articles that explain the GRADE system for rating the quality of evidence and strength of recommendations

Testing makes a variety of contributions to patient care

Clinicians use tests—including signs and symptoms, imaging, and biochemistry—to identify physiological derangements, establish prognosis, monitor illness, and diagnose.[1] This article focuses on diagnosis: the use of tests to establish the presence or absence of a disease (such as tuberculosis), target condition (such as iron deficiency), or syndrome (such as Cushing's syndrome).

Clinicians often use diagnostic tests as a package or strategy. For example, in managing patients with apparently operable lung cancer, clinicians may proceed directly to thoracotomy or apply a strategy of imaging the brain, bone, liver, and adrenal glands, with subsequent management depending on the results. Thus, one can often think of evaluating or recommending not a single test, but a diagnostic strategy. Guideline panels considering a diagnostic test or strategy should begin by identifying the patients, diagnostic intervention (strategy), comparison, and outcomes of interest (box).[2 3]

Test accuracy is a surrogate for outcomes important to patients

The main contribution of this article is that it presents a framework for thinking about the quality of evidence for diagnostic tests in terms of their impact on outcomes important to patients ("patient-important outcomes"). Usually, when clinicians think about diagnostic tests, they focus on accuracy (sensitivity and specificity); that is, how well the test classifies patients correctly as having or not having a disease. The underlying assumption is, however, that obtaining a better idea of whether a target condition is present or absent will result in improved outcome. For patients who present with apparently operable lung cancer, the presumption is that additional tests will spare patients the morbidity and early mortality associated with futile thoracotomy. The example of computed tomography for coronary artery disease in the box illustrates another common rationale for a new test: replacement of another test (coronary computed tomography instead of conventional angiography) to avoid complications associated with a more invasive and expensive alternative.[6]

The best way to assess any diagnostic strategy—but in particular new strategies with putative superior accuracy—is a randomised controlled trial in which investigators randomise patients to experimental or control diagnostic approaches and measure mortality, morbidity, symptoms, and quality of life (figure).[7]

When diagnostic intervention studies—ideally randomised controlled trials but also observational studies—comparing the impact of alternative diagnostic

Table 1 | Examples and implications of different testing scenarios

Example of new test and reference test or strategy	Putative benefit of new test	Diagnostic accuracy	
		Sensitivity	Specificity
Shorter version of dementia test compared with original mini mental state exam for diagnosis of dementia	Simpler test, less time	Equal	Equal
Helical computed tomography for renal calculus compared with intravenous pyelogram (IVP)	Detection of more (but smaller) calculi	Greater	Equal
Computed tomography for coronary artery disease compared with coronary angiography	Less invasive testing, but misses some cases	Slightly less	Less

See text for explanations of terms.

Example of a sensible clinical question

In patients in whom coronary artery disease is suspected, does multislice spiral computed tomography of coronary arteries as a replacement for conventional invasive coronary angiography reduce complications with acceptable rates of false negatives associated with coronary events and false positives leading to unnecessary treatment and complications?[4 5]

strategies on patient-important outcomes are available, guideline panels can use the GRADE approach described in previous articles in this series.[12 13] When such studies are not available, guideline panels must focus on studies of test accuracy and make inferences about the likely impact on patient-important outcomes.[14] The key questions are whether a reduction in false negatives (cases missed) or false positives and corresponding increases in true positives and true negatives will occur, how accurately similar or different patients are classified by the alternative testing strategies, and what outcomes occur in both patients labelled as cases and those labelled as not having disease. Table 1 presents examples that illustrate these questions.

Using indirect evidence to make inferences about impact on patient-important outcomes

Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes requires the availability of effective treatment.[1] Alternatively, even without an effective treatment, an accurate test may be beneficial if it reduces test related adverse effects or anxiety, or if confirming a diagnosis improves patients' well-being through the prognostic information it imparts.

For instance, the results of genetic testing for Huntington's chorea, an untreatable condition, may provide either welcome reassurance that a patient will not have the condition or the ability to plan for the future knowing that he or she will develop the condition. The ability to plan is analogous to an effective treatment, and the benefits of planning need to be balanced against the downsides of receiving an early diagnosis.[15-17] We will now describe factors that influence the balance between desirable and undesirable consequences, focusing on the quality of evidence. We will use a simplified approach that classifies test results into yielding true positives, true negatives, false positives, and false negatives.

Victor M Montori associate professor, Knowledge and Encounter Research Unit, Department of Medicine, Mayo Clinic College of Medicine, Rochester, MN 55905, USA

Patrick Bossuyt professor, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam 1100 DE, Netherlands

Gordon H Guyatt professor, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

For the GRADE Working Group

Correspondence to:
schuneh@mcmaster.ca

Judgment about quality of underlying evidence

Study design and limitations (risk of bias)

GRADE's four categories of quality of evidence represent a gradient of confidence in estimates of the effect of a diagnostic test strategy on patient-important outcomes.[13] Table 2 describes how GRADE deals with the particular challenges of judging the quality of evidence of alternative diagnostic strategies. As we have noted, randomised trials of alternative diagnostic approaches represent the ideal study design for informing recommendations. Nevertheless, in the GRADE system, valid studies of test accuracy also start as high quality in the diagnostic framework. Such studies are, however, vulnerable to limitations and often provide low quality evidence for recommendations as a result of the indirect evidence they usually offer on impact on patient-important outcomes.

Valid studies of diagnostic test accuracy include representative and consecutive patients in whom legitimate diagnostic uncertainty exists—that is, the sort of patients to whom clinicians would apply the test in the course of regular clinical practice. If studies fail this criterion—and, for example, enrol severe cases and healthy

focusing on accuracy

Patients' outcomes and expected impact on management				Balance between presumed outcomes, test complications, and cost
True positives	True negatives	False positives	False negatives	
Presumed influence on patient-important outcomes:				Evidence of shorter time and similar test accuracy (and thus patients' outcomes) would generally support new test's usefulness
Uncertain benefit from earlier diagnosis and treatment	Almost certain benefit from reassurance	Likely anxiety and possible morbidity from additional testing and treatment	Possible detriment from delayed diagnosis	
Directness of evidence (test results) for outcomes important to patients:				
Some uncertainty	No uncertainty	Some uncertainty	Major uncertainty	
Presumed influence on patient-important outcomes:				Fewer complications and downsides compared with IVP would support new test's usefulness, but balance between desirable and undesirable effects is not clear in view of uncertain consequences of identifying smaller stones
Certain benefit for larger stones; less clear benefit for smaller stones, and unnecessary treatment can result	Almost certain benefit from avoiding unnecessary tests	Likely detriment from unnecessary additional invasive tests	Likely detriment for large stones; less certain for small stones, but possible detriment from unnecessary additional invasive tests for other potential causes of complaints	
Directness of evidence (test results) for patient-important outcomes:				
Some uncertainty	No uncertainty	No uncertainty	Major uncertainty	
Presumed influence on patient-important outcomes:				Undesirable consequences of more false positives and false negatives with computed tomography are not acceptable despite higher rate of rare complications (infarction and death) and higher cost of angiography
Benefit from treatment and fewer complications	Benefit from reassurance and fewer complications	Harm from unnecessary treatment	Detriment from delayed diagnosis or myocardial insult	
Directness of evidence (test results) for patient-important outcomes:				
No uncertainty	No uncertainty	No uncertainty	Some uncertainty	

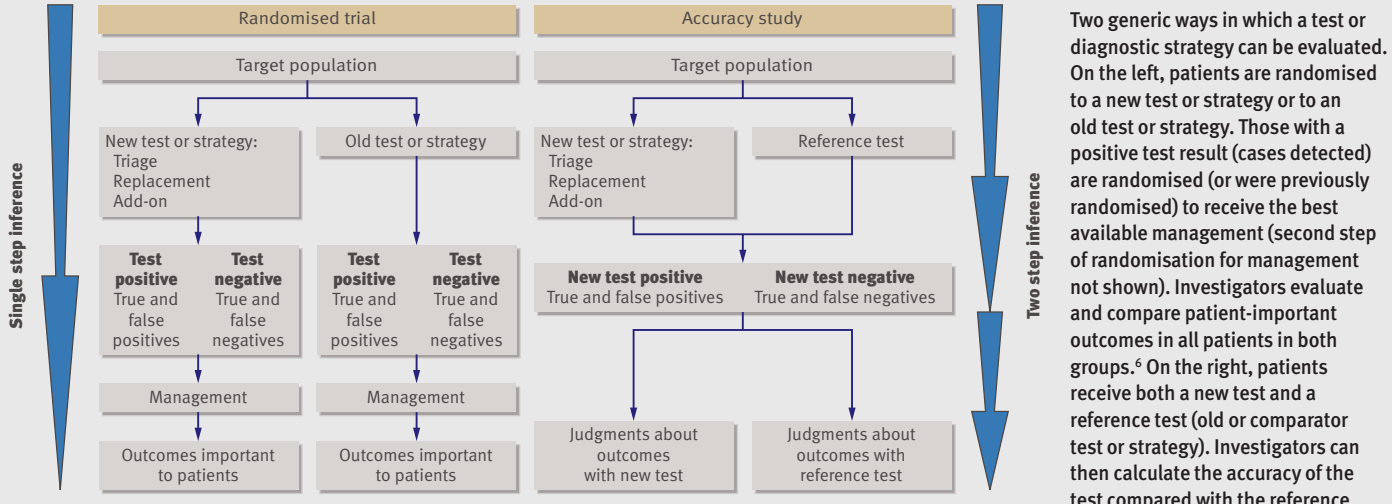
controls—the apparent accuracy of a test is likely to be misleadingly high.[18 19] Valid studies involve a comparison between the test or tests under consideration and an appropriate reference (sometimes called "gold") standard. Investigators' failure to make such a comparison in all patients increases the risk of bias. The risk of bias is further increased if the people who carry out or interpret the test are aware of the results of the reference or gold standard test or vice versa. Guideline panels can use existing instruments to assess the risk of bias in studies evaluating the accuracy of diagnostic tests and can downgrade the quality of evidence if serious limitations exist.[20-22]

Directness

Judging directness presents perhaps the greatest challenges for guideline panels making recommendations about diagnostic tests. For instance, a new test may be simpler to do, with lower risk and cost, but may produce false positives and false negatives. Consider the consequences of replacing invasive angiography with coronary computed tomography scanning for the diagnosis of

Table 2 | **Factors that decrease quality of evidence for studies of diagnostic accuracy and how they differ from evidence for other interventions**

Factors that determine and can decrease quality of evidence	Explanations and differences from quality of evidence for other interventions
Study design	Different criteria for accuracy studies—Cross sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are considered high quality and can move to moderate, low, or very low depending on other factors
Limitations (risk of bias)	Different criteria for accuracy studies—Consecutive patients should be recruited as a single cohort and not classified by disease state, and selection as well as referral process should be clearly described.[7] Tests should be done in all patients in the same patient population for new test and well described reference standard; evaluators should be blind to results of alternative test and reference standard
Indirectness:	
Outcomes	Similar criteria—Panels assessing diagnostic tests often face an absence of direct evidence about impact on patient-important outcomes. They must make deductions from studies of diagnostic tests about the balance between the presumed influences on patient-important outcomes of any differences in true and false positives and true and false negatives in relation to complications and costs of the test. Therefore, accuracy studies typically provide low quality evidence for making recommendations owing to indirectness of the outcomes, similar to surrogate outcomes for treatments
Patient populations, diagnostic test, comparison test, and indirect comparisons	Similar criteria—Quality of evidence can be reduced if important differences exist between populations studied and those for whom recommendation is intended (in previous testing, spectrum of disease or comorbidity); if important differences exist in tests studied and diagnostic expertise of people applying them in studies compared with settings for which recommendations are intended; or if tests being compared are each compared with a reference (gold) standard in different studies and not directly compared in same studies
Important inconsistency in study results	Similar criteria—For accuracy studies, unexplained inconsistency in sensitivity, specificity, or likelihood ratios (rather than relative risk or mean differences) can reduce quality of evidence
Imprecise evidence	Similar criteria—For accuracy studies, wide confidence intervals for estimates of test accuracy or true and false positive and negative rates can reduce quality of evidence
High probability of publication bias	Similar criteria—High risk of publication bias (for example, evidence from small studies for new intervention or test, or asymmetry in funnel plot) can lower quality of evidence



Example
Randomised control trials (RCTs) explored a diagnostic strategy guided by the use of B type natriuretic peptide (BNP)—designed to aid diagnosis of heart failure—compared with no use of BNP in patients presenting to the emergency department with acute dyspnoea.[8 9] As it turned out, the group randomised to receive BNP spent a shorter time in the hospital at lower cost, with no increased mortality or morbidity

Example
Consistent evidence from well designed studies shows fewer false negative results with non-contrast helical computed tomography (CT) than with intravenous pyelography (IVP) in the diagnosis of suspected acute urolithiasis.[10] However, the stones in the ureter that CT detects but IVP "misses" are smaller, and hence are likely to pass more easily. As RCTs evaluating the outcomes in patients treated for smaller stones are not available, the extent to which reduction in cases that are missed (false negatives) and follow-up of incidental findings unrelated to renal calculi with CT have important health benefits remains uncertain[11]

Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, patients are randomised to a new test or strategy or to an old test or strategy. Those with a positive test result (cases detected) are randomised (or were previously randomised) to receive the best available management (second step of randomisation for management not shown). Investigators evaluate and compare patient-important outcomes in all patients in both groups.[6] On the right, patients receive both a new test and a reference test (old or comparator test or strategy). Investigators can then calculate the accuracy of the test compared with the reference test (first step). To make judgments about importance to patients of this information, patients with a positive test (or strategy) in either group are (or have been in previous studies) submitted to treatment or no treatment; investigators then evaluate and compare patient-important outcomes in all patients in both groups (second step)

Table 3 | Key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography* be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?[5]

Measure	Test findings (95% CI)
Pooled sensitivity	0.96 (0.94 to 0.98)
Pooled specificity	0.74 (0.065 to 0.84)
Positive likelihood ratio†	5.4 (3.4 to 8.3)
Negative likelihood ratio†	0.05 (0.03 to 0.09)

*Assuming that the reference standard, angiography, does not yield false positives or false negatives.

†Average likelihood ratios from Hamon et al.[5]

coronary artery disease (tables 3 and 4). True positive results will lead to the administration of treatments of known effectiveness (drugs, angioplasty and stents, bypass surgery), and true negative results will spare patients the possible adverse effects of the reference standard test. On the other hand, false positive results will result in adverse effects (unnecessary drugs and interventions, including the possibility of follow-up angioplasty) without apparent benefit, and false negatives will result in patients not receiving the benefits of available interventions that help to reduce the subsequent risk of coronary events.

Thus, it is relatively certain that minimising false positives and false negatives will benefit patients. The impact of inconclusive test results is less clear, but they are clearly undesirable. Furthermore, the complications of invasive angiography—infarction and death—although rare, are undoubtedly important. When guideline panels balance the desirable and undesirable consequences of diagnostics tests, they should consider the importance of these consequences for patients. In this example of patients with a relatively low probability for coronary artery disease, computed tomography scanning results in a large number of false positives leading to unnecessary anxiety and further testing (table 4). It also leads to missing about 1% (false negatives) of patients who have coronary artery disease.

Guideline panels considering questions of diagnosis also face the same sort of challenges regarding indirectness as do panels making recommendations for other interventions.[2] Test accuracy may vary across populations of patients, so panels need to consider how well the populations included in the studies correspond to the population that is the focus of the recommendations. Similarly, panels need to consider how comparable new tests and reference tests are to the tests used in the settings for which the recommendations are made. Finally, when evaluating two or more alternative new tests or strategies, panels need to consider whether these diagnostic strategies were compared directly (in one study) or indirectly (in separate studies) with a common (reference) standard.[25-27]

Arriving at a bottom line for study quality

Table 5 shows the evidence summary and the quality assessment for all critical outcomes of computed tomography angiography as a replacement for invasive angiography. Little or no uncertainty exists about the directness of the evidence (for test results) for patient-important outcomes for true positives, false positives, and true negatives (table 1). However, some uncertainty about the extent to which limitations in test accuracy will have deleterious consequences on patient-important outcomes for false negatives led to downgrading the quality of evidence from high to moderate (table 5 see bmj.com). Unexplained heterogeneity in the results across studies further reduced the quality of evidence for all outcomes. Major uncertainty about the impact of false negative tests on patient-important outcomes would have led to downgrading the quality of evidence from high to low for the other examples in table 1.

Arriving at a recommendation

The balance of presumed patient-important outcomes as the result of true and false positives and negatives with test complications determine whether a guideline panel makes a recommendation for or against applying a test.[12] Other factors influencing the strength of a recommendation include the quality of the evidence, the uncertainty about values and preferences associated with the tests and presumed patient-important outcomes, and cost.

Coronary computed tomography scanning avoids the adverse consequences of invasive angiography, which can include myocardial infarction and death. These consequences are, however, very rare. As a result, a guideline panel evaluating coronary computed tomography as a replacement test for coronary angiography could, despite its lower cost, make a weak recommendation against its use in place of invasive coronary angiography. This recommendation follows from the large number of false positives and the risk of

Table 4 | Consequences of key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography* be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?[6]

Consequences	No per 1000 patients	Importance†
True positive results‡	192	8
True negative results§	592	8
False positive results¶	208	7
False negative results**	8	9
Inconclusive results††§§	–	5
Complications‡‡§§	–	5
Cost[§§]	–	5

All results given per 1000 patients tested for prevalence of 20% and likelihood ratios shown in table 3.

*Assuming that the reference standard, angiography, does not yield false positives or false negatives.

†On a 9 point scale, GRADE recommends classifying these outcomes as not important (score 1-3), important (4-6), and critical (7-9) to a decision.[13 18 19]

‡Important because mandates drugs, angioplasty and stents, bypass surgery.

§Important because spares patients unnecessary interventions associated with adverse effects.

¶Important because patients are exposed to unnecessary potential adverse effects from drugs and invasive procedures.

**Important because increase risk of coronary events as a result of patients not receiving efficacious treatment.

††Uninterpretable, indeterminate, or intermediate test results; important because generate anxiety, uncertainty as to how to proceed, further testing, and possible negative consequences of either treating or not treating.

‡‡Not reliably reported; important because although rare, they can be serious.

§§Although the data for these consequences are not reported for simplicity or because they are not exactly known on the basis of the available data, they are important.

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

missing patients with coronary artery disease who could be treated effectively (false negatives). It also follows from the evidence for the new test being only low quality and the consideration of values. Despite the general preference for less invasive tests with lower risks of complications, most patients would probably favour the more invasive approach (angiography), given the risks associated with false positives and negatives.

Conclusion

As for other management recommendations, the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests provides a comprehensive and transparent approach for developing these recommendations. Recognising that test results are surrogates for patient-important outcomes is central to this approach. The application of the approach requires a shift in clinicians' thinking to clearly recognise that, whatever their accuracy, diagnostic tests are of value only if they result in improved outcomes for patients.

We thank the many people and organisations that have contributed to the progress of the GRADE approach through funding of meetings and feedback on the work described in this article.

The members of the Grade Working Group are Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Robin Harbour, Margaret Haugh, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Nicola Magrini, Merce Marzo, James Mason, Jacek Mrukowics, Andrew D Oxman, Susan Norris, Vivian Robinson, Holger J Schünemann, Jane Thomas, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, and James Woodcock.

Contributors: All listed authors, and other members of the GRADE working group, contributed to the development of the ideas in the manuscript, and read and approved the manuscript. HJS wrote the first draft and collated comments from authors and reviewers for subsequent iterations. All other listed authors contributed ideas about structure and content and provided feedback. HJS is the guarantor.

Funding: This work was partially funded by "The human factor, mobility and Marie Curie Actions Scientist Reintegration" European Commission Grant: IGR 42192-"GRADE" to HJS.

Competing interests: The authors are members of the GRADE Working Group. The work with this group probably advanced the careers of some or all of the authors and group members. Authors listed in the byline have received travel reimbursement and honorariums for presentations that included a review of GRADE's approach to grading the quality of evidence and strength of recommendations. GHG acts as a consultant to UpToDate; his work includes helping UpToDate in their use of GRADE. HJS is documents editor and methodologist for the American Thoracic Society; one of his roles in these positions is helping implement the use of GRADE; he supports the implementation of GRADE by organisations worldwide. VMM supports the implementation of GRADE in several North American not for profit professional organisations.

1 Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
3 Mulrow C, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;4:288-95.
4 Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club* 2004;140(1):A11-2.
5 Hamon M, Biondi-Zoccai GG, Malagutti P, Agostoni P, Morello R, Valgimigli M, et al. Diagnostic performance of multislice spiral computed tomography of coronary arteries as compared with conventional invasive coronary angiography: a meta-analysis. *J Am Coll Cardiol* 2006;48:1896-910.
6 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92.
7 Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356:1844-7.
8 Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004;350:647-54.
9 Moe G, Howlett J, Januzzi JL, Zowall H, Canadian multicenter improved management of patients with congestive heart failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. *Circulation* 2007;115:3103-10.
10 Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40:280-6.
11 Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53:144-8.
12 Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ. Going from evidence to recommendations. *BMJ* 2008, doi: 10.1136/bmj.39493.646875.AE.
13 Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008, doi: 10.1136/bmj.39490.551019.BE.
14 Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;144:850-5.
15 Maat-Kievit A, Vegter-van der Vlis M, Zoeteweij M, Losekoot M, van Haeringen A, Roos R. Paradox of a better test for Huntington's disease. *J Neurol Neurosurg Psychiatry* 2000;69:579-83.
16 Walker FO. Huntington's disease. *Semin Neurol* 2007;27:143-50.
17 Almqvist EW, Brinkman RR, Wiggins S, Hayden MR. Psychological consequences and predictors of adverse events in the first 5 years after predictive testing for Huntington's disease. *Clin Genet* 2003;64:300-9.
18 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
19 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
20 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138:40-4.
21 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
22 Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.
23 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
24 Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
25 Fletcher RH. Carcinoembryonic antigen. *Ann Intern Med* 1986;104:66-73.
26 Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. *Am J Med* 1984;77:64-71.
27 Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* 1990;81:815-20.

CRITICAL APPRAISAL OF A DIAGNOSTIC STUDY
DIAGNOSIS WORKSHEET

Citation:

Are the results of this diagnostic study valid?	
1. Was there a clearly defined, focused research question? What was the study question?	
2. Was the index test described in sufficient detail to permit its replication?	
3. Was the presence or absence of the target condition confirmed with a validated, appropriate "gold" or reference standard? In other words, is the reference standard likely to correctly classify the target condition? Was the reference standard described in sufficient detail to permit its replication?	
4. Was the diagnostic test evaluated in an appropriate spectrum of patients? Was the spectrum of patients representative of the patients who will receive the test in practice? Or was it evaluated among patients with confirmed/severe disease and compared against healthy volunteers (case-control approach)?	
5. Did the whole sample or a random selection of the sample, receive verification using the reference standard test?	
6. Was the reference standard applied to all patients, regardless of the index test result? Did the results of the test being evaluated influence the decision to perform the reference standard?	
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	
8. Were the index test results interpreted without knowledge of (i.e. blinded) the results of the reference standard?	
9. Were the reference standard results interpreted without knowledge of (i.e. blinded) the results of the index test?	
10. Were uninterpretable/indeterminate test results accounted for and reported?	
11. Were withdrawals from the study accounted for?	

Any other potential biases in this study?

What are the results?

SAMPLE CALCULATIONS:

		Target Disorder (iron deficiency anaemia)				Totals	
		Present		Absent			
Diagnostic Test Result (serum ferritin)	Positive (<65 mmol/L)	731	a	b	270	a+b	1001
	Negative (≥65 mmol/L)	78	c	d	1500	c+d	1578
	Totals	809	a+c	b+d	1770	a+b+c+d 2579	

Sensitivity = a/(a+c) = 731/809 = 90%

Specificity = d/(b+d) = 1500/1770 = 85%

Likelihood Ratio for a positive test result = LR+=sens/(1-spec)=90%/15%=6

Likelihood Ratio for a negative test result=LR-=(1-sens)/spec=10%/85%=0.12

Positive Predictive Value = a/(a+b) = 731/1001 = 73%

Negative Predictive Value = d/(c+d) = 1500/1578 = 95%

YOUR CALCULATIONS:

		Target Disorder			Totals
		Present		Absent	
Diagnostic Test Result	Positive		a	b	a+b
	Negative		c	d	c+d
	Totals		a+c	b+d	a+b+c+d

Sensitivity = a/(a+c) =

Specificity = d/(b+d) =

Likelihood Ratio for a positive test result = LR+=sens/(1-spec)=

Likelihood Ratio for a negative test result=LR-=(1-sens)/spec=

Positive Predictive Value = a/(a+b) =

Negative Predictive Value = d/(c+d) =

Are the likelihood ratios likely to be useful in routine clinical practice?

Can you apply the results to patient care?

1. Is the diagnostic test available, affordable, and feasible in your setting?	
2. Will the results be applicable to the patients in your setting? Is your patient so different from those in the study that its results can't help you?	
3. Can you generate a clinically sensible estimate of your patient's pre-test probability (from practice data, from personal experience, from the report itself, or from clinical speculation)	
4. Will the resulting post-test probabilities change your management strategy and help your patient? (Could it move you across a test-treatment threshold?)	
5. Would the consequences of the test help your patient? (Will patients be better off as a result of the test?)	

In summary:

What are the major strengths of this study?
What are the major limitations of this study?
Are there any major ethical concerns with this study?

APPENDIX:

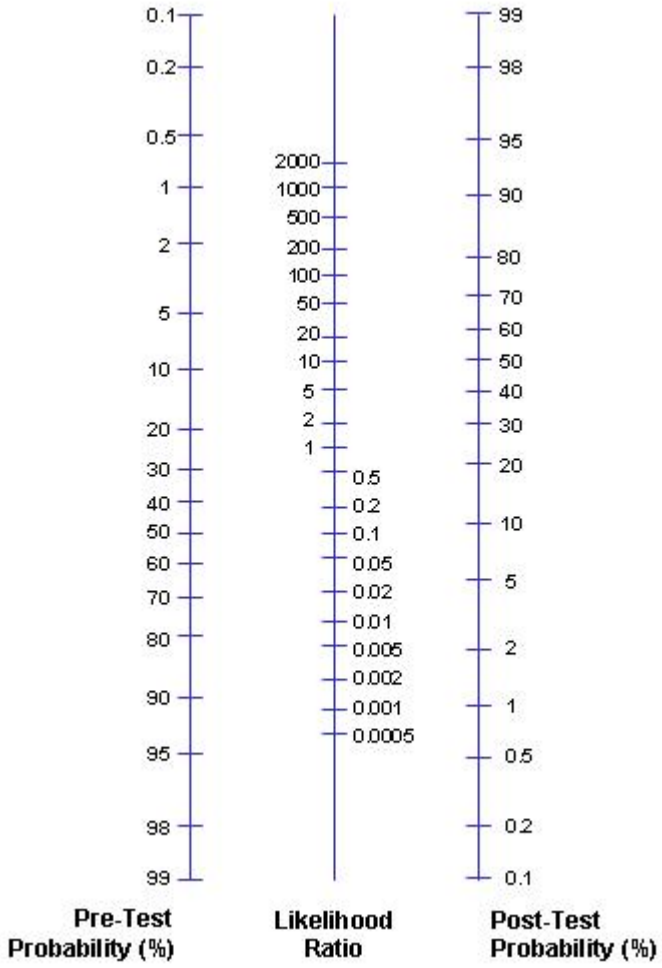
Likelihood Ratio Nomogram

The Likelihood Ratio (LR) is the likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that that same result would be expected in a patient without the target disorder. The LR is used to assess how good a diagnostic test is and to help in selecting an appropriate diagnostic test(s) or sequence of tests. They have advantages over sensitivity and specificity because they are less likely to change with the prevalence of the disorder, they can be calculated for several levels of the symptom/sign or test, they can be used to combine the results of multiple diagnostic test and the can be used to calculate post-test probability for a target disorder.

Formula for computing post-test probability, given pre-test probability and LR:

Post-test odds = pre-test odds X LR
Pre-test odds = pre-test probability / (1-pre-test probability)
Post-test probability = post-test odds / (post test odds+1)

To make it easier to move between odds and probability, the LR nomogram below can be used:



A LR greater than 1 produces a post-test probability which is higher than the pre-test probability. An LR less than 1 produces a post-test probability which is lower than the pre-test probability. When the pre-test probability lies between 30 and 70 per cent, test results with a very high LR (say, above 10) rule in disease. An LR below 1 produces a post-test probability less than the pre-test probability. A very low LR (say, below 0.1) virtually rules out the chance that the patient has the disease.

QUADAS Tool for Quality Assessment of Diagnostic Studies*

Citation:

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
2.	Were selection criteria clearly described?	( )	( )	( )
3.	Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	( )	( )	( )
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
6.	Did patients receive the same reference standard regardless of the index test result?	( )	( )	( )
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	( )	( )	( )
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
13.	Were uninterpretable/ intermediate test results reported?	( )	( )	( )
14.	Were withdrawals from the study explained?	( )	( )	( )

Based on the QUADAS checklist, how do you rate the quality of this study?

NOTES