Evaluation of Tuberculosis Diagnostics: Establishing an Evidence Base Around the Public Health Impact

Richard J. Lessells,[1,2] Graham S. Cooke,[2,3] Marie-Louise Newell,[3,4] and Peter Godfrey-Faussett[1]

[1]Department of Clinical Research, London School of Hygiene and Tropical Medicine, and [2]Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Mtubatuba, KwaZulu-Natal, South Africa; [3]UCL Institute of Child Health, and [4]Department of Infectious Diseases, Imperial College, London, United Kingdom

The limitations of existing tuberculosis diagnostic tools are significantly hampering tuberculosis control efforts, most noticeably in areas with high prevalence of human immunodeficiency virus (HIV) infection and antituberculosis drug resistance. However, renewed global interest in tuberculosis research has begun to bear fruit, with several new diagnostic technologies progressing through the development pipeline. There are significant challenges in building a sound evidence base to inform public health policies because most diagnostic research focuses on the accuracy of individual tests, with often significant limitations in the design, conduct, and reporting of diagnostic accuracy studies. Diagnostic accuracy studies may not be appropriate to guide public health policies, and clinical trials may increasingly be required to determine the incremental value and cost-effectiveness of new tools. The urgent need for new diagnostics should not distract from pursuing rigorous scientific evaluation focused on public health impact.

Global control of the tuberculosis epidemic is a public health priority [1, 2]. The targets for reduction in tuberculosis prevalence and mortality linked to the Millennium Development Goals and enshrined in the STOP TB Global Plan 2006–2015 will not be achieved with current interventions [3, 4]. There is an acute need for improved tuberculosis diagnostics as one critical component of the public health response to the tuberculosis epidemic.

The rapid growth of the human immunodeficiency virus (HIV) epidemic and the emergence of antituberculosis drug resistance have highlighted the major deficiencies in current diagnostic technologies both for pathogen detection and for diagnosis of drug resistance [5]. In most high-burden countries, sputum smear microscopy remains the principal tool for diagnosing active disease; however, operationally, its sensitivity for pulmonary tuberculosis can be as low as 20% [6, 7]. Sputum culture and drug susceptibility testing are available in certain settings, but their impact is limited by the long duration and complexity of the laboratory processes [8]. Additional challenges are faced in developing diagnostics for extrapulmonary tuberculosis, pediatric tuberculosis, and latent tuberculosis infection [9–11].

The STOP TB Global Plan 2006–2015 included the target that, "by 2010, simple, robust, affordable technologies for use at peripheral levels of the health system will enable rapid, sensitive detection of active tuberculosis at the first point of care" [4, p. 24]. Although this has not been achieved, there have been developments in the tuberculosis diagnostic field, and promising technologies have entered the clinical sphere [6, 12–15]. Most promising has been the Xpert MTB/RIF system, an automated molecular test that simultaneously detects *Mycobacterium tuberculosis* and mutations associated with rifampicin resistance [16, 17]. It is hoped that the

Correspondence: Richard J. Lessells, MBChB, Department of Clinical Research, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, United Kingdom (richard.lessells@lshtm.ac.uk).

The Journal of Infectious Diseases 2011;204:S1187–95

0022-1899 (print)/1537-6613 (online)/2011/204S4-0011$14.00
DOI: 10.1093/infdis/jir412

renewed global focus on tuberculosis will in the next few years lead to the further proliferation of diagnostic technologies in parallel with advances in therapeutics and vaccines.

It is the responsibility of the global scientific community to correctly evaluate these new technologies so that proven effective and cost-effective diagnostics can be adopted, thus generating the greatest public health impact. The importance of diagnostic research in the overall tuberculosis research agenda has been highlighted by many different groups [2, 15, 18–22]. However, huge gaps in funding for tuberculosis research and tuberculosis control remain [1, 2, 23]; this should force us to rethink how diagnostic research can be most effectively targeted and rationalized to inform public health policies.

This article focuses on the framework for evaluation of new diagnostics: at the outset, we look at the potential benefits of new diagnostics, and then we discuss different methodologies to evaluate diagnostic performance with a view to their ultimate implementation. Our focus throughout is on diagnostic tests for detection of active tuberculosis disease and/or drug resistance in high-burden countries.

POTENTIAL IMPACT OF NEW TUBERCULOSIS DIAGNOSTICS

It has been hypothesized that a test more sensitive than sputum microscopy for tuberculosis would be the diagnostic intervention that would alleviate the greatest burden of infectious disease in developing countries [24]. More specifically, one mathematical model of the global tuberculosis epidemic suggested that a new rapid diagnostic test with 100% sensitivity, 100% specificity, and 100% access could prevent 625 000 deaths annually (equivalent to 36% of all tuberculosis-related deaths) [25]. Other models have derived fairly consistent estimates of mortality reductions of 17%–23% from a more sensitive rapid tuberculosis diagnostic, despite exploring different epidemics [26–28]. In one model, the estimated benefit in terms of mortality from a new diagnostic test was equivalent in magnitude to that expected from a novel vaccine or an optimized 2-month treatment regimen for active disease [26]. This highlights 2 important points: (1) no single intervention will have the impact required to meet tuberculosis control targets; thus, scaled-up investment in research and implementation of diagnostics, drugs, and vaccines will be required; and (2) because new diagnostics could have an equivalent impact to new drugs or vaccines, evaluation of diagnostics should be as rigorous as evaluation of drugs and vaccines.

EXISTING FRAMEWORK FOR TUBERCULOSIS DIAGNOSTIC RESEARCH AND DEVELOPMENT

The fact that sputum smear microscopy remains the cornerstone of tuberculosis diagnosis in most high-burden countries is

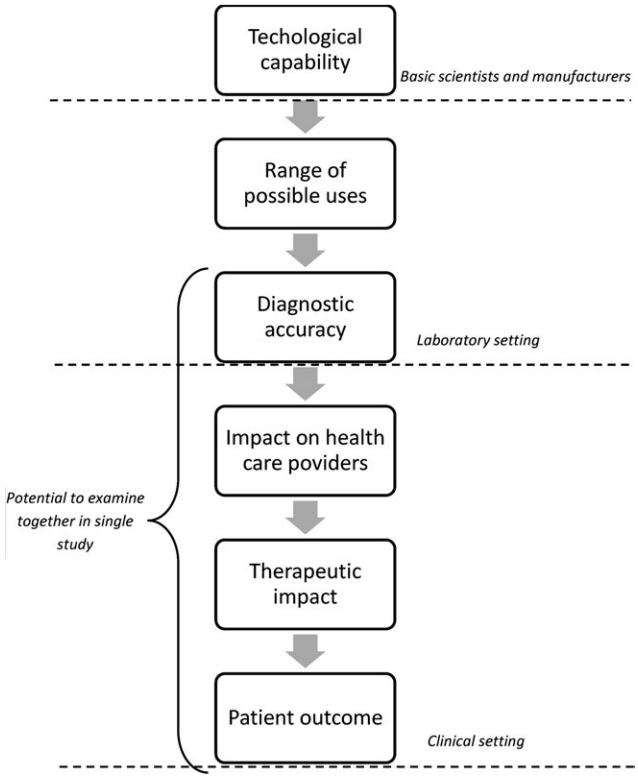

Figure 1. Stepwise approach to evaluation of diagnostic technologies.

testament to the relative paucity of research and development in the diagnostic arena and the failure to translate research findings into policy. In medicine broadly, diagnostic research tends to be performed in stepwise fashion, with basic science leading to laboratory-based performance evaluation and then to clinical studies (Figure 1) [29]. This structure inherently tends to exclude the perspectives of end users in the conception and development of diagnostics, although more recently in the tuberculosis field, organizations have assisted this process by defining the ideal specifications for a point-of-care test [30].

In the tuberculosis field, the process of diagnostic development has rarely gone beyond diagnostic accuracy studies to assess the impact in clinical practice on clinical decision making, patient outcomes, and health system costs [13, 31, 32]. This is in part explained by the fact that the regulatory framework for in vitro diagnostic devices usually does not require evidence beyond performance data. Diagnostic accuracy studies are an important part of the evaluation process. However, there is much potential for bias in such studies, and diagnostic accuracy might vary widely between different clinical settings and populations [33–36].

In the field of diagnostic accuracy research, there have been certain key initiatives aimed at improving and standardizing research methodologies and reporting: the guidelines for diagnostic evaluation produced by the TDR Diagnostics Evaluation Expert Panel (DEEP) [37], the Quality Assessment of Diagnostic

Item	
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?

Figure 2. The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool.

Accuracy Studies (QUADAS) tool [38], and the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) initiative [39, 40]. The DEEP guidelines outline best practice in the design and conduct of diagnostic evaluations, with focus on performance characteristics and operational feasibility. QUADAS is a quality assessment tool to be used specifically for the assessment of diagnostic accuracy studies included in systematic reviews. The tool consists of 14 items (Figure 2); the majority involve sources of bias, with a few relating to variability and quality of reporting. The objective of the STARD initiative is to improve the quality of reporting of diagnostic accuracy studies. The 25-item checklist (Figure 3) allows the reader to judge the potential for bias (internal validity) and the generalizability and applicability (external validity) of the study.

A systematic review that used both QUADAS and STARD criteria to assess tuberculosis diagnostic accuracy studies published during 2004–2006 showed significant deficiencies in methodology and reporting of studies [41]. Unfortunately, more widespread use of the STARD system has not been apparent in recent years. As a further example, of the 10 published studies evaluating the diagnostic accuracy of the Genotype MTBDR*plus* assay (published during 2007–2010) [42–51], only one manuscript explicitly mentions STARD [51]. Additional efforts are required by researchers, research funders, journal editors, and policy makers to encourage the use of these tools, with the aim of improving the quality and validity of this element of the evidence base.

THE NEED FOR HIGH-QUALITY EVIDENCE TO INFORM PUBLIC HEALTH POLICIES

Public health policies and guidelines are now usually informed by a systematic approach to judging the relevant evidence. In the tuberculosis field, the World Health Organization (WHO) convenes expert groups to assess the available evidence for a specific intervention (eg, diagnostic test), and this group then presents their findings to the WHO Strategic and Technical Advisory Group for Tuberculosis (STAG-TB) for consideration and endorsement. The system to assess the evidence now adopted by many organizations, including WHO, is the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system, which incorporates judgments on the quality of evidence (high, moderate, low, or very low) and on the strength of any recommendation (initially categorized as strong or weak; now incorporates "conditional," whereby national programs should consider implementation based on their own situation) [52, 53].

The GRADE system is based around the concept of patient-important outcomes, and as such, evidence from diagnostic interventions creates additional challenges. Studies using indirect outcomes (eg, diagnostic accuracy studies) will usually provide lower-quality evidence because of the uncertainty about outcomes important to patients and the potential for bias [54]. It is important to be clear that the rating of low quality in this context does not necessarily imply that studies were conducted poorly, but that data from the study are not optimal for deriving public health recommendations.

GOING BEYOND DIAGNOSTIC ACCURACY STUDIES—THE NEED FOR IMPACT DATA

In the STOP TB New Diagnostics Working Group blueprint for the evaluation of diagnostics, the next step after diagnostic accuracy studies are demonstration studies, which include patient outcomes (Figure 4) [55]. These demonstration studies are designed to assess the scaled-up test performance and to determine patient-level outcomes. This is the stage of the evaluation process that should start to inform policy. It is stated in this document that patient-important outcomes should be assessed (eg, time to initiation of treatment, time to smear and/or culture conversion, and treatment outcome) and

Section	Item	
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		
Participants	3	The study population: 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, 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 (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		
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 treatment, 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 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

Figure 3. Standards for the Reporting of Diagnostic Accuracy Studies (STARD) checklist.

that "these impact-related data should be compared to historical data recorded prior to implementation of the new test in routine clinical practice" [55, p. 62]. This use of historical data is problematic as a method of assessing any health care intervention and would not generally be accepted by regulatory bodies in the field of drugs or vaccines [56]. It is difficult to be

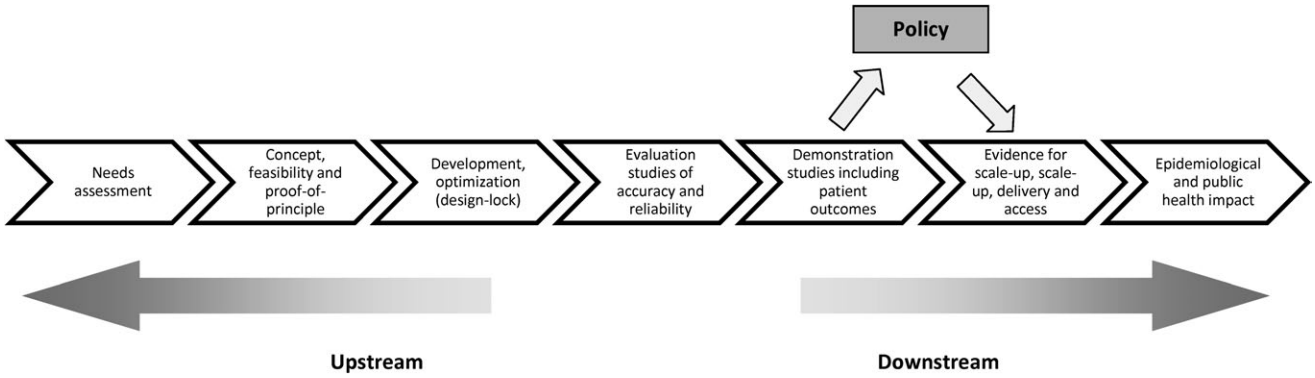

Figure 4. The pathway for evaluation of new diagnostics (from the STOP TB New Diagnostic Working Group).

sure that any comparison is fair; there are potential sources of bias, and consequently, the risk is that the value of the intervention can be exaggerated.

Two organizations that have been instrumental in driving forward development and evaluation of diagnostic technologies for tuberculosis are the Foundation for Innovative New Diagnostics and the WHO TDR program (Special Programme for Research and Training in Tropical Diseases). Demonstration studies are key elements of their tuberculosis projects, which aim to determine the feasibility, impact, and cost-effectiveness of the diagnostic test under evaluation. The evidence from these studies is a key element assessed by the expert groups and reported to STAG-TB. If we take the example of the Genotype MTBDR*plus* assay, preliminary data regarding patient-important outcomes from the South African demonstration projects seemed relatively disappointing because the median turnaround times did not meet their predefined objective of 7 days; of the patients with multidrug-resistant tuberculosis who were identified, only 28% were started on appropriate therapy on the basis of the test result (42% had therapy delayed until results of conventional drug susceptibility testing were available) [57]. Although these results were based only on preliminary data analysis and are understandable during implementation of a new technology, there has, to our knowledge, been no further published evidence from high-burden settings on patient-important outcomes. However, the test has been introduced into routine practice in some countries, and its use is now being scaled up [58].

It is generally considered that the optimal methodology for assessing the clinical impact of any intervention, including diagnostics, is the randomized controlled trial (RCT) [59–61]. This is the methodology least prone to bias in estimating the benefits and risks of any intervention. Data from RCTs can additionally be used to perform economic evaluation, a step of major importance for policy makers. The relative shortage of RCTs in diagnostic research, in contrast to therapeutic and vaccine research, is likely to be explained by a combination of factors: lack of emphasis on this level of evidence by manufacturers and regulatory authorities, limited funding and poor coordination of diagnostic research, and logistical and ethical challenges. There are features specific to diagnostic trials that complicate trial design and implementation. In a tuberculosis diagnostic study, the population of interest might be persons with suspected pulmonary tuberculosis (eg, individuals with cough). Inevitably, the majority of participants will not have tuberculosis; thus, the potential effect size on the total cohort resulting from improved diagnosis is relatively small. However, we have to include the entire cohort in a trial if we want to capture comprehensive outcome data (to balance benefits and harms).

To reveal the value of well-designed RCTs in diagnostic research, it is worthwhile to stop studying tuberculosis and consider malaria, another global health priority. Malaria rapid diagnostic tests (RDTs) have been shown to have good diagnostic accuracy [62], and mathematical models have suggested that implementation of RDTs could lead to significant public health benefits in settings where malaria is endemic [63]. Trials were designed to assess the performance of the tests in a field setting and to measure the impact on health care providers, therapeutic decisions, and patient outcomes [64–67]. Three of these trials showed that, despite good diagnostic accuracy, there was no reduction in incorrect antimalarial treatment with the use of RDTs [64–66]; of more concern, one trial even showed a significant reduction in correct antimalarial treatment [66]. These trials have provided vital information for the further development and implementation of RDTs. The results of these trials highlight the fact that a diagnostic test is only ever a vehicle to guide therapies; it is never of therapeutic benefit, and it is the treatment decision that will impact on patient outcomes.

CONCEPTUALIZING CLINICAL TRIALS OF TUBERCULOSIS DIAGNOSTICS

The first step in any trial is to determine the hypothesis that is to be tested because this will inform the trial design. It is important to consider the likely position of the new test in the diagnostic

process. In the case of a test for active pulmonary tuberculosis, we need to decide how the test will be introduced in the existing diagnostic structure, which includes sputum microscopy, sputum culture, drug-susceptibility testing, and chest radiography. It could be proposed as a replacement for ≥1 of these tests, as an addition to these tests, or as a means of triage, for example, to target sputum culture and/or drug-susceptibility testing. This decision is in turn likely to depend on the proposed benefits of the new test (eg, whether it is more rapid, more sensitive, more specific, less technical, safer, or less expensive). Furthermore, we need to consider the outcomes of interest, whether related to benefit or harm; these may be appropriate or inappropriate commencement of tuberculosis treatment, outcomes during treatment (smear or culture conversion), final treatment outcomes (cure or completion), and mortality.

One possible reason to explain the lack of RCTs in diagnostic research is the perception that diagnostic tests carry minimal or no risk. Although the test is unlikely to harm the patient, the consequences of the test (eg, the therapeutic decision) may confer harm, as shown in the example of RDTs of malaria. What risks might we expect in a trial of a tuberculosis diagnostic? Consider a hypothetical trial comparing clinical outcomes between a rapid molecular tuberculosis test and the standard-of-care diagnostic pathway (Figure 5). At a basic level, this trial will tell us whether the benefits from earlier correct diagnosis or exclusion of tuberculosis outweigh the risks from incorrect classification of disease (false-negative or false-positive results). The benefits would seem to be self-evident but need to be quantified. The risks are more complicated and will be context specific. False-negative diagnoses will result in appropriate treatment being withheld, with potential for poorer outcomes. False-positive diagnoses also carry risk, however, because alternative diagnoses may not be considered and, therefore, not treated, and patients may be exposed to potentially toxic therapy. For diagnosis of drug resistance, the risks from incorrect classification are even more complicated. False-negative results of genotypic testing may lead to inappropriate treatment with first-line regimens, with consequent adverse outcomes, including amplification of drug resistance. False-positive results may lead to inappropriate treatment with multidrug-resistant tuberculosis regimens, with lower efficacy against sensitive strains and with risks of severe toxicity.

These examples highlight another challenge with tuberculosis diagnostic research (and common to much diagnostic research), which is the lack of a perfect gold standard with which to compare new tests. If our new test is potentially more sensitive than the existing test (as might be the case with molecular tests, compared with sputum culture), this will affect any analysis. The lack of a gold standard often requires a construct gold standard that comprises information from the reference test with additional clinical information and follow-up information [68]. Of further concern, discrepancies between

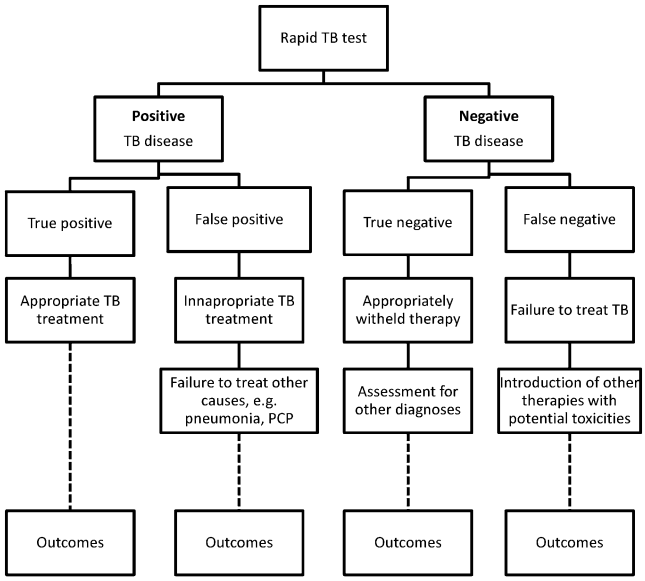

Figure 5. Potential impact of false-positive and false-negative tuberculosis diagnoses in a hypothetical trial comparing a rapid molecular test to tuberculosis culture.

phenotypic and genotypic drug-susceptibility results can be extremely difficult to interpret, and it is not always clear which is the more reliable measure of drug resistance [69]. In many ways, these issues reinforce the need for well-designed clinical trials because thorough interpretation of the tests may only be possible with meticulously collected baseline and follow-up clinical data.

PRACTICAL TRIAL DESIGNS

If the outcomes of interest are individual-level outcomes (eg, treatment initiation and mortality), a clinical trial with individual randomization would be the logical and statistically most efficient design. However, because there will be information regarding the diagnostic performance from the laboratory-based evaluation, the question arises, if the test is shown to have comparable accuracy to an existing test but has other advantages (ie, more rapid and/or less invasive), is it ethical to conduct an RCT with individual randomization? Critical to this decision is whether there is equipoise regarding the clinical outcome. Equipoise with regard to clinical outcomes of a diagnostic strategy arises, for example, when the consequences of misdiagnosis are severe (eg, HIV-infected patients who receive a misdiagnosis of tuberculosis who are dying of another HIV-related illness) or when failure to diagnose does not lead to mistreatment or poorer outcomes (eg, patients prescribed tuberculosis treatment regardless of the test result).

Individual randomization may, however, present considerable logistical challenges in certain health care settings, and for this reason, cluster randomized designs may be considered with

health care units (eg, hospitals, clinics, and mobile teams) as clusters. Cluster randomized designs are increasingly used in public health research. The principal reasons for considering such a design are as follows: if the intervention is to be delivered to groups rather than individuals, if the outcome is to be measured at a population level, or to avoid contamination by individuals in the same community who are randomized to different trial arms [70]. However, there is also an acceptance that cluster randomization may also be appropriate in settings where it offers greater logistical convenience, compared with an individually randomized trial, although cluster RCTs generally require larger sample sizes and have added challenges in design, analysis, and ethics [70–72].

A further modification of the cluster randomized design is the phased implementation or stepped-wedge design [70, 73]. The key features of this design are that all clusters receive the intervention by the end of the trial, and the order in which the clusters receive the intervention is decided at random. This is particularly appropriate when there is preexisting evidence that the intervention may have a beneficial effect and when assigning clusters to the control arm for the duration of the trial might be ethically unacceptable. This might be particularly suited to evaluation of certain diagnostic technologies, for which there is evidence from initial diagnostic accuracy studies that suggests beneficial effect.

If randomization is not deemed to be appropriate or feasible, alternative prospective trial designs, often termed quasi-experimental designs, may still be able to generate evidence on the effectiveness of diagnostics [74]. An example would be the pre- and postimplementation study in which outcomes are measured during a pre-intervention phase and subsequently during a postintervention phase. Although the lack of randomization threatens the internal validity (no firm conclusion can be made with regard to the effect of the intervention unless the effect size is large), there may conversely be a gain in external validity (improved generalizability of findings if fewer patients are excluded than in conventional RCTs).

Retrospective studies may be the only methodology to obtain outcome data in circumstances in which a diagnostic is widely implemented on the basis of performance characteristics. Such pre- and postimplementation analyses have been used in high-resource settings to estimate the impact of molecular resistance testing on detection and treatment of multidrug-resistant tuberculosis [75, 76].

Whether a clinical trial is justified in the evaluation of diagnostics will ultimately depend on the balance between the benefit to be gained by accurately establishing the impact of a new tool and the costs of running a large clinical trial and potentially delaying full-scale implementation of an effective intervention. These decisions are not straightforward, and collaboration between scientists and policy makers is vital to determine when diagnostic trials are necessary.

CONCLUSIONS

Recent developments in tuberculosis diagnostics have led to much optimism, but we still lack the tools that meet the needs of patients in high-burden countries. The next 10–20 years will hopefully see further developments in diagnostic technology. We need to ensure that the framework for evaluating diagnostic tools is best suited to ensuring that the tools with the greatest public health impact and cost-effectiveness are implemented and that those with minimal impact are developed further or are discarded. Diagnostic accuracy studies are an important early step in the evaluation process but do not produce sufficient evidence to inform public health policies. Well-designed prospective studies (including RCTs) should be integrated in the research pathway to provide reliable information on therapeutic impact, patient outcomes, and cost-effectiveness. This new era of tuberculosis diagnostics should be accompanied by a new era for diagnostic research focused clearly on the evaluation of public health impact.

Notes

Financial Support. This work was supported by the Wellcome Trust (grant 090999).

Potential conflicts of interest. All authors: no reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. World Health Organization. Global tuberculosis control: epidemiology, strategy, financing: WHO report 2009. Geneva, Switzerland: World Health Organization, **2009**.
2. Marais BJ, Raviglione MC, Donald PR, et al. Scale-up of services and research priorities for diagnosis, management, and control of tuberculosis: a call to action. Lancet **2010**; 375:2179–91.
3. United Nations General Assembly. Road map toward the implementation of the United Nations Millennium Declaration. New York: United Nations, **2002**.
4. Stop TB Partnership/World Health Organization. The global plan to stop TB, 2006–2015. Geneva, Switzerland: World Health Organization, **2006**.
5. Abdool Karim SS, Churchyard GJ, Abdool Karim Q, Lawn SD. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. Lancet **2009**; 374:921–33.
6. 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.
7. Steingart KR, Ramsay A, Pai M. Optimizing sputum smear microscopy for the diagnosis of pulmonary tuberculosis. Expert Rev Anti Infect Ther **2007**; 5:327–31.
8. Urbanczik R, Rieder HL. Scaling up tuberculosis culture services: a precautionary note. Int J Tuberc Lung Dis **2009**; 13:799–800.
9. World Health Organization. Improving the diagnosis and treatment of smear-negative pulmonary and extrapulmonary tuberculosis among adults and adolescents. Recommendations for HIV-prevalent and resource-constrained settings. Geneva, Switzerland: World Health Organization, **2007**.
10. Shingadia D, Novelli V. Diagnosis and treatment of tuberculosis in children. Lancet Infect Dis **2003**; 3:624–32.

11. 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**; 146:340–54.
12. Pai M, Kalantri S, Dheda K. New tools and emerging technologies for the diagnosis of tuberculosis: part II. Active tuberculosis and drug resistance. Expert Rev Mol Diagn **2006**; 6:423–32.
13. 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.
14. Grandjean L, Moore DA. Tuberculosis in the developing world: recent advances in diagnosis with special consideration of extensively drug-resistant tuberculosis. Curr Opin Infect Dis **2008**; 21:454–61.
15. World Health Organization/Stop TB Partnership. The global plan to stop TB 2011–2015: transforming the fight towards elimination of tuberculosis. Geneva, Switzerland: World Health Organization, **2010**.
16. Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med **2010**; 363:1005–15.
17. Helb D, Jones M, Story E, et al. Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. J Clin Microbiol **2010**; 48:229–37.
18. Chaisson RE, Harrington M. How research can help control tuberculosis. Int J Tuberc Lung Dis **2009**; 13:558–68.
19. Cobelens FG, Heldal E, Kimerling ME, et al. Scaling up programmatic management of drug-resistant tuberculosis: a prioritized research agenda. PLoS Med **2008**; 5:e150.
20. Fauci AS. Multidrug-resistant and extensively drug-resistant tuberculosis: the National Institute of Allergy and Infectious Diseases Research agenda and recommendations for priority research. J Infect Dis **2008**; 197:1493–8.
21. Wallis RS, Pai M, Menzies D, et al. Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. Lancet **2010**; 375:1920–37.
22. 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.
23. Treatment Action Group. 2009 report on tuberculosis research funding trends, 2005–2008. 2nd ed. New York: Treatment Action Group, 2010.
24. Mabey D, Peeling RW, Ustianowski A, Perkins MD. Diagnostics for the developing world. Nat Rev Microbiol **2004**; 2:231–40.
25. 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.
26. 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.
27. 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 U S A **2008**; 105:11293–8.
28. 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.
29. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ **1986**; 134:587–94.
30. Medicins sans Frontieres/Treatment Action Group/ Partners in Health. Specifications for point-of-care TB tests: expert opinion check from TB field practitioners. http://www.msfaccess.org/TB_POC_Parismeeting/fileadmin/user_upload/diseases/tuberculosis/TB%20POC%20Full%20-Survey%20Analysis%20report.pdf. Accessed 13 October 2010.
31. Small PM, Perkins MD. More rigour needed in trials of new diagnostic agents for tuberculosis. Lancet **2000**; 356:1048–9.
32. Pai M, O'Brien R. Tuberculosis diagnostics trials: do they lack methodological rigor? Expert Rev Mol Diagn **2006**; 6:509–14.
33. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med **1978**; 299: 926–30.
34. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA **1999**; 282:1061–6.
35. 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.
36. 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.
37. The TDR Diagnostics Evaluation Expert Panel. Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol **2008**; 6:S16–S28.
38. 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.
39. 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**; 138:40–4.
40. 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:W1–12.
41. Fontela PS, Pant Pai N, 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.
42. Albert H, Bwanga F, Mukkada S, et al. Rapid screening of MDR-TB using molecular Line Probe Assay is feasible in Uganda. BMC Infect Dis **2010**; 10:41.
43. Anek-Vorapong R, Sinthuwattanawibool C, Podewils LJ, et al. Validation of the GenoType MTBDR*plus* assay for detection of MDR-TB in a public health laboratory in Thailand. BMC Infect Dis **2010**; 10:123.
44. Barnard M, Albert H, Coetzee G, O'Brien R, Bosman ME. Rapid molecular screening for multidrug-resistant tuberculosis in a high-volume public health laboratory in South Africa. Am J Respir Crit Care Med **2008**; 177:787–92.
45. Bazira J, Asiimwe BB, Joloba ML, Bwanga F, Matee MI. Use of the GenoType(R) MTBDR*plus* assay to assess drug resistance of Mycobacterium tuberculosis isolates from patients in rural Uganda. BMC Clin Pathol **2010**; 10:5.
46. Evans J, Stead MC, Nicol MP, Segal H. Rapid genotypic assays to identify drug-resistant Mycobacterium tuberculosis in South Africa. J Antimicrob Chemother **2009**; 63:11–6.
47. Hillemann D, Rusch-Gerdes S, Richter E. Evaluation of the GenoType MTBDR*plus* assay for rifampin and isoniazid susceptibility testing of Mycobacterium tuberculosis strains and clinical specimens. J Clin Microbiol **2007**; 45:2635–40.
48. Huang WL, Chen HY, Kuo YM, Jou R. Performance assessment of the GenoType MTBDR*plus* test and DNA sequencing in detection of multidrug-resistant Mycobacterium tuberculosis. J Clin Microbiol **2009**; 47:2520–4.
49. Huyen MN, Tiemersma EW, Lan NT, et al. Validation of the GenoType MTBDR*plus* assay for diagnosis of multidrug resistant tuberculosis in South Vietnam. BMC Infect Dis **2010**; 10:149.
50. Matsoso LG, Veriava Y, Poswa X, et al. Validation of a rapid tuberculosis PCR assay for detection of MDR-TB patients in Gauteng, South Africa. South Afr J Epidemiol Infect **2010**; 25:12–5.
51. Nikolayevskyy V, Balabanova Y, Simak T, Malomanova N, Fedorin I, Drobniewski F. Performance of the Genotype MTBDR*plus* assay in the diagnosis of tuberculosis and drug resistance in Samara, Russian Federation. BMC Clin Pathol **2009**; 9:2.
52. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ **2004**; 328:1490.
53. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ **2008**; 336:924–6.

54. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ **2008**; 336:1106–0.
55. 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.
56. Pocock SJ. Justification for randomized controlled trials. In: Clinical trials: a practical approach. Chichester, UK: John Wiley & Sons Ltd., **1983**:50–65.
57. World Health Organization. Molecular line proble assays for rapid screening of patients at risk of multi-drug resistant tuberculosis (MDR-TB). Expert Group Report. Geneva, Switzerland: World Health Organization, **2008**.
58. World Health Organization and UNITAID. Expanding and accelerating access to diagnostics for patients at risk of multidrug-resistant tuberculosis. http://www.who.int/tb/publications/factsheet_expand_tb.pdf. Accessed 13 October 2010.
59. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ **2002**; 324:477–80.
60. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet **2000**; 356:1844–7.
61. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol **2009**; 62:364–73.
62. World Health Organization. Malaria rapid diagnostic test performance: results of WHO product testing of malaria RDTs: round 1 (2008). Geneva, Switzerland: World Health Organization, **2009**.
63. Rafael ME, Taylor T, Magill A, Lim YW, Girosi F, Allan R. Reducing the burden of childhood malaria in Africa: the role of improved diagnostics. Nature **2006**; 444(Suppl 1):39–48.
64. Reyburn H, Mbakilwa H, Mwangi R, et al. Rapid diagnostic tests compared with malaria microscopy for guiding outpatient treatment of febrile illness in Tanzania: randomised trial. BMJ **2007**; 334:403.
65. Bisoffi Z, Sirima BS, Angheben A, et al. Rapid malaria diagnostic tests vs. clinical management of malaria in rural Burkina Faso: safety and effect on clinical decisions. A randomized trial. Trop Med Int Health **2009**; 14:491–8.
66. Skarbinski J, Ouma PO, Causer LM, et al. Effect of malaria rapid diagnostic tests on the management of uncomplicated malaria with artemether-lumefantrine in Kenya: a cluster randomized trial. Am J Trop Med Hyg **2009**; 80:919–26.
67. Ansah EK, Narh-Bana S, Epokor M, et al. Rapid testing for malaria in settings where microscopy is available and peripheral clinics where only presumptive treatment is available: a randomised controlled trial in Ghana. BMJ **2010**; 340:c930.
68. 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:iii–ix–51.
69. Van Deun A, Barrera L, Bastian I, et al. Mycobacterium tuberculosis strains with highly discordant rifampin susceptibility test results. J Clin Microbiol **2009**; 47:3501–6.
70. Hayes RJ, Moulton LH. Cluster randomised trials. Interdisciplinary Statistics Series. Boca Raton, FL: CRC Press, **2009**.
71. Edwards SJ, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. BMJ **1999**; 318:1407–9.
72. Osrin D, Azad K, Fernandez A, et al. Ethical challenges in cluster randomized controlled trials: experiences from public health interventions in Africa and Asia. Bull World Health Organ **2009**; 87:772–9.
73. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol **2006**; 6:54.
74. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company, **2002**.
75. Banerjee R, Allen J, Lin SY, et al. Rapid drug susceptibility testing with a molecular beacon assay is associated with earlier diagnosis and treatment of multidrug-resistant tuberculosis in California. J Clin Microbiol **2010**; 48:3779–81.
76. O'Riordan P, Schwab U, Logan S, et al. Rapid molecular detection of rifampicin resistance facilitates early diagnosis and treatment of multi-drug resistant tuberculosis: case control study. PLoS One **2008**; 3:e3173.