Published in final edited form as:
Lancet Infect Dis. 2010 June ; 10(6): 387–394. doi:10.1016/S1473-3099(10)70071-2.

Sputum Monitoring during Tuberculosis Treatment for Predicting Outcome: A Systematic Review and Meta-analysis

David J Horne, MD, MPH,
Division of Pulmonary and Critical Care Medicine, University of Washington School of Medicine, 325 9th Ave, Box 359762, Seattle, WA 98104, USA

Sarah E Royce, MD, MPH,
Division of Pulmonary and Critical Care Medicine & Global Health Sciences, (both) University of California, San Francisco, San Francisco, CA

Lisa Gooze, MD,
TB Control Program, San Mateo County Public Health, 225 West 37th Ave, San Mateo, CA 94403, USA

Masa Narita, MD,
Public Health - Seattle & King County, Tuberculosis Control Program, Division of Pulmonary and Critical Care Medicine, University of Washington School of Medicine, 325 9th Ave, Box 359776, Seattle, WA 98104, USA

Philip C Hopewell, MD,
Division of Pulmonary and Critical Care Medicine & Francis J. Curry National Tuberculosis Center, (both) University of California, San Francisco, 1001 Potrero Avenue, 5K1, San Francisco, CA 94110, USA

Payam Nahid, MD, MPH, and
Division of Pulmonary and Critical Care Medicine & Francis J. Curry National Tuberculosis Center, (both) University of California, San Francisco, 1001 Potrero Avenue, 5K1, San Francisco, CA 94110, USA

Karen R Steingart, MD, MPH
Francis J. Curry National Tuberculosis Center, University of California, San Francisco, San Francisco, CA, USA

Summary

Corresponding Author: David J Horne, MD, Division of Pulmonary and Critical Care Medicine, University of Washington School of Medicine, 325 9th Ave, Box 359762, Seattle, WA 98104, USA, dhorne@u.washington.edu, Telephone: 206-372-4372, Fax: 206-744-8584.

Conflicts of Interest
We declare that we have no conflicts of interest.

Author Contributions
Horne and Steingart had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Gooze, Hopewell, Horne, Royce, Steingart
Screening of studies: Gooze, Horne, Steingart
Analysis and interpretation of data: Hopewell, Horne, Royce, Steingart
Drafting of the manuscript: Horne, Royce, Steingart
Critical revision of the manuscript for important intellectual content: Gooze, Hopewell, Horne, Nahid, Narita, Royce, Steingart

As this was an analysis of published data, no ethics committee approval was sought.

In new pulmonary tuberculosis (TB) patients, the World Health Organization previously recommended performing sputum smear examination at the end of the second month of therapy and, if positive, to extend the intensive phase. We performed a systematic review and meta-analysis to evaluate the accuracy of a positive sputum smear or culture during treatment for predicting failure or relapse in pulmonary TB. We searched PubMed, EMBASE, and the Cochrane Library for studies published in English through December 2009. We included randomized controlled trials, cohort, and case-control studies of previously untreated pulmonary TB patients who had received a standardized regimen with rifampin in the initial phase. Accuracy results were summarized in forest plots and pooled using a hierarchical regression approach. Fifteen papers met inclusion criteria. The pooled sensitivities for both the 2 month smear (24%, 95% CI 12–42, 6 studies) and culture (40%, 95% CI 25–56, 4 studies) to predict relapse were low. Corresponding specificities (85%, 95% CI 72–90) and (85%, 95% CI 77–91) were higher, but modest. For failure, the 2 month smear (7 studies) had low sensitivity (57%, 95% CI 41–73) and higher, though modest, specificity (81%, 95% CI 72–87). Both sputum smear microscopy and mycobacterial culture during TB treatment have low sensitivity and modest specificity for predicting failure and relapse. Although we pooled a diverse group of patients, the individual studies had similar performance characteristics. Better predictive markers are needed.

INTRODUCTION

Tuberculosis (TB) is a major global health problem with nine million new cases and almost two million deaths per year.[1] The World Health Organization (WHO) recommends that patients with previously-untreated pulmonary TB receive a four-drug regimen during the two-month initial phase of treatment that includes rifampin.[2] The overall rate of failure or relapse (poor outcome) in patients receiving directly observed treatment, short-course (DOTS) with a rifampin-containing regimen is low.[3–5] In patients receiving six-month regimens who have drug-susceptible organisms, the estimated failure rate is 1–4% and the relapse rate 7% or less.[3] Relapse of TB continues to place a significant burden on the patient and TB control programs; worldwide in 2007, at least 270,000 patients returned after relapse (5% of TB notifications).[1] Early identification of patients who have an increased risk of a poor outcome coupled with an intervention, such as treatment modification, could potentially reduce this burden.

In the past, treatment guidelines have recommended the examination of a sputum smear at the end of the two-month initial phase of treatment and, if the smear is positive, to continue the initial phase for an extra month before proceeding with the standard four or six month continuation phase.[2] These recommendations, designed for low resource settings that lack the ability to perform culture or drug susceptibility testing, have been questioned due to lack of evidence, and the ability of a positive sputum examination result to predict poor outcomes has not been fully assessed.[6–11] A recently published review on TB biomarkers identified an inverse relationship between two-month sputum culture conversion and relapse, although the utility of two-month culture conversion as a guide to treatment of individual patients was limited by poor positive predictive value.[12] However, this review was not a systematic review of all available evidence.

To estimate the accuracy of a positive sputum smear and/or culture for predicting poor outcome in patients with pulmonary TB who received a standardized regimen that included rifampin in the initial phase, we performed a systematic review and meta-analysis of the literature. In addition to estimating the accuracy of the sputum examination result, the quality of studies was appraised. Results from this systematic review were used in the development of the most recent WHO Treatment of tuberculosis guidelines.[13]

STUDY POPULATION AND METHODS

We used standard methods for systematic reviews of diagnostic accuracy studies.[14–17]

Search strategy and selection criteria

We searched PubMed, EMBASE, and the Cochrane Library for studies published in the English language through May 2008. We updated the literature search through December 31, 2009 with no language restrictions. Although we did not limit the earliest date of publication, this was practically limited by the earliest clinical investigation of rifampin in 1965.[18] Our search terms included *tuberculosi*s or *Mycobacterium tuberculosis* and *sputum/microbiology* or *sputum/cytology* or *recurrence* or *treatment failure* or *relapse* (Appendix). Bibliographies of original articles and reviews were reviewed for additional relevant studies.

The study selection process is shown in Figure 1. We included randomized control trials, cohort and case-control studies that met the following selection criteria: (1) Participants were previously untreated patients with pulmonary TB diagnosed by either a positive sputum smear or culture; (2) Participants received standardized (not individualized) treatment with rifampin in at least the initial phase of treatment; (3) Sputum smear or culture examination was performed during treatment; (4) Outcomes were treatment failure and relapse. The following studies were excluded: (1) studies that exclusively used rifabutin or rifapentine; (2) individual arms of randomized controlled trials if they involved patients who received a non-rifampin containing regimen; (3) studies involving patients who had only extrapulmonary TB (i.e. no pulmonary involvement); (4) non-human studies; (5) studies that exclusively enrolled patients undergoing re-treatment of TB; (6) studies involving children (less than 12 years of age), as they are more likely to have paucibacillary disease; (7) studies where data on outcome by sputum examination result were unavailable (i.e. for 2 × 2 tables for true positives, false positives, false negatives, and true negatives).[19] When possible, we excluded study arms that included only retreatment patients.[20] Three authors (LG, DJH, KRS) were responsible for review of titles and/or abstracts. At least two reviewers independently screened each study using pre-specified inclusion and exclusion criteria (see Figure 1). Disagreements on study selection were resolved by consensus. Three hundred ninety-one publications were retrieved for full text review.

Data extraction and quality assessment

We created and piloted a data extraction form with a subset of eligible studies. Based on the experience gained in the pilot, the data extraction form was revised and finalized. Two reviewers independently abstracted data from the included studies with the standardized form on the following characteristics: study design, population, methodology, geographic area, results of sputum assessment, outcomes, treatment regimen, and supervision of treatment. Additional data were requested from authors as needed.

The quality of studies was appraised using a subset of criteria from QUADAS, a validated tool for diagnostic studies, as well as additional criteria.[21] These criteria, summarized in Table 1, included study design, manner of patient selection, loss of patients during treatment and follow-up, duration of follow-up, type of follow-up (i.e. active, passive) supervision of treatment, and use of culture confirmation.[22] A single publication could contribute multiple studies if different outcomes, or months or types of sputum examination were reported. At the end of the selection process, 15 papers (28 studies) were included in the analysis.

Data synthesis and analysis

We used WHO definitions for TB outcomes.[2] Specifically, treatment failure is defined as positive sputum smear microscopy (or culture) at the 5th month or later during TB

treatment. Relapse is defined as a patient who previously completed TB treatment successfully and subsequently is again found to have bacteriologically positive (sputum smear or culture) pulmonary TB. All included studies reported at least 12 months of follow-up after completion of treatment.

Values for sensitivity and specificity were calculated for each study, along with their 95% confidence intervals. Sensitivity refers to the proportion of TB patients with a poor outcome (relapse or failure) who had a positive sputum smear or culture at a given month; specificity refers to the proportion of TB patients not experiencing a poor outcome who were sputum negative at a given month. We used forest plots to summarize results according to type of sputum specimen, month of specimen examination, and outcome. We derived odds ratios and pooled estimates of sensitivity and specificity using hierarchical summary receiver operating characteristic (HSROC) analysis.[23] The advantage of HSROC analysis is that it jointly models sensitivity and specificity, weights studies according to the number of participants, and takes into account unmeasured heterogeneity among studies by using random effects.[15] The odds ratio is the odds of a positive result in individuals experiencing a poor outcome compared to the odds of a positive result in individuals that did not experience a poor outcome and is a global measure of test performance.[24] We performed the HSROC analyses in Stata IC/10.0 (Stata Corporation, USA) with the user written command "metandi".[23, 25] If fewer than four studies were available, their estimates were pooled by means of a fixed effects model using Meta-DiSc software (version 1.4, Madrid, Spain because HSROC random effects models do not converge. [26]

Predictive values were determined in STATA/IC 10.0. In this review, positive predictive value (PPV) is the proportion of those with a positive sputum test result who fail or relapse, and can be interpreted as the probability that a positive result is correct. Negative predictive value (NPV) is the proportion of those with a negative sputum test result who do not fail or relapse, and can be interpreted as the probability that a negative result is correct.

Heterogeneity refers to the degree of variability in accuracy estimates across studies. It is a concern in meta-analyses because if significant heterogeneity is present then summary estimates are not meaningful. Heterogeneity may be due to variability in test thresholds, prevalence of treatment failure or relapse, the populations studied, and reference standard tests used.[27] Heterogeneity was assessed by visual examination of forest plots of sensitivities and specificities and by Chi-squared and I-squared tests using Meta-DiSc software.[26] We recognized that studies were heterogeneous in many respects, particularly concerning the type of sputum specimen (smear or culture) examined during treatment, month of specimen examination, outcome (relapse or failure), and manner (active or passive) of patient follow-up. Based on our research questions, we pre-specified subgroups by type and month of sputum examination, and outcome (Figure 2). We also evaluated the impact of manner of patient follow-up on our findings.

RESULTS

Characteristics of included studies

The literature searches identified over 12,000 citations from which 15 publications (28 studies) were selected (Figure 1).[6, 8, 11, 20, 28–38] Table 1 summarizes study quality; Tables 2a and b show characteristics of individual studies by relapse (18 studies) or treatment failure (10 studies), respectively. Twenty-three (82%) studies were conducted in low-income countries. The majority of studies assessed the examination of a sputum specimen at month two, consistent with standard practice.[2] The total number of participants was 34,575; median 346 (interquartile range 229 to 418).

As seen in Table 1, most studies used a cohort study design and provided supervised treatment. Fifteen (54%) studies used culture to confirm TB diagnosis. In 11 (39%) studies the poor outcome was confirmed by culture; in 13 (46%) studies confirmed by smear, and in four studies, by either smear or culture. Eleven (39%) studies reported losing fewer than 10% of participants during treatment. Among the 18 studies evaluating relapse, a majority conducted active follow-up for relapse. Eleven (61%) studies reported losing fewer than 10% of participants during a 12-month follow-up period after treatment completion. Most studies did not report the number of sputum specimens obtained during treatment.

Three studies included retreatment patients; when reported, retreatment patients comprised less than 20% of the patients in each study.[6, 28, 31] Two studies specifically excluded HIV-infected patients.[8, 28] Two studies reported enrolling HIV-infected individuals, who were 28% and 5% of the total enrolled patients, respectively.[29, 30] The remaining papers did not assess HIV-status, although in one study the HIV co-infection rate among TB patients in the study country was reported to be approximately 35%.[35] A result from drug susceptibility testing was reported in twenty studies, where any drug resistance was 8% or less except for the study by Santha that reported 16% prevalence of resistance to one or more drugs.[6, 8, 28, 29, 31, 32, 34, 37, 38]

Accuracy for predicting outcomes

Figures 3 and 4 show sensitivity and specificity estimates for all sputum specimen analyses displayed in forest plots. For both relapse and failure, sensitivity estimates were low across studies except for the study by Ramarokoto, which was unusual in that treatment was extended if subjects remained smear-positive at 2 months.[33] However, we included this study in our meta-analyses as exclusion of this study did not significantly change our results. Specificity values across studies were modest.

Tables 2a and 2b show the accuracy in each study of either culture or smear to predict a poor outcome by the month of specimen examination. Pooled estimates for sensitivity and specificity and odds ratios are shown in Table 3. For culture predicting relapse, culture at month two (4 cohort studies) had low pooled sensitivity, 40% (95% CI 25–56), moderate pooled specificity, 85% (95% CI 77–91) and an odds ratio of 3.8 (95% CI 2.2–6.8).[8, 29, 31, 38] Chi-squared and I-squared tests suggested moderate heterogeneity across studies for sensitivity (Chi-squared=5.47; 3 degrees of freedom [p =0.14]; inconsistency [I-squared] =45.1%). We investigated a possible source of heterogeneity by evaluating a subset of studies (3 studies) that reported actively following patients after treatment completion.[8, 31, 38] Compared with the sensitivity for all studies using 2 month culture to predict relapse, studies with active follow-up showed similar sensitivity [51% (95% CI 40–63)] and less variability (Chi-squared=0.95; degrees of freedom=2 [p=0.62]; inconsistency [I-squared]=0.0 %), suggesting differences in the manner of follow-up contributed to heterogeneity in the results.

For predicting relapse, compared with culture at month two, smear at month two (6 studies) yielded lower pooled sensitivity 24% (95% CI 12–42), with similar pooled specificity 83% (95% CI 72–90) (Table 3).[8, 20, 31–34] The pooled odds ratio was 1.5 (95% CI 1.1–2.2). Sensitivity for heterogeneity was high (Chi-squared =13.29; degrees of freedom=5 [p=0.02]; inconsistency [I-squared] =62.4%). For predicting failure, smear at month two (7 studies) yielded low sensitivity 57% (95% CI 41–73) and modest specificity 81% (95% CI 72–87); [8, 11, 33–37] The pooled odds ratios was 5.8 (95% CI 4.3–7.8). Again, there was substantial heterogeneity for sensitivity across studies, (Chi-squared =22.31; degrees of freedom=6 [p=0.001]; inconsistency [I-squared] =73.1%), making meaningful interpretation difficult.

Both culture and smear had low PPVs (9 to 18%) in predicting a poor outcome, suggesting a low probability that a positive sputum specimen at any month could correctly predict failure or relapse (Table 3). In contrast, NPVs were high (at least 93%), indicating a negative sputum test result at any month during treatment makes relapse or failure unlikely.

DISCUSSION

We performed a systematic review and meta-analysis of the accuracy of sputum examination during treatment to identify pulmonary TB patients who will fail treatment or experience relapse. We found low sensitivity and moderate specificity for prediction of relapse or failure in all studies regardless of sputum examination or time of evaluation. These results were similar for individual studies and in the pooled analyses.

When an individual's status is unknown, predictive values are used to estimate the probability that the outcome will occur based on a test result.[39, 40] Assuming that the incidence of relapse and failure is 7% and 3%, respectively, we found poor PPVs and good NPVs for both individual studies (data not presented) and pooled studies. The low PPVs indicate that a positive sputum result during treatment does not imply that an individual will experience a poor outcome from TB. High NPVs indicate that a negative smear or culture examination during treatment implies that an individual will be unlikely to experience treatment failure or relapse from TB.

Until recently, the WHO recommended extension of the intensive phase of treatment in patients with positive sputum smears at the end of the second month of TB treatment.[2] Informed, in part, by the present systematic review and others, the recently completed fourth edition of the WHO Treatment of Tuberculosis Guidelines no longer carries this recommendation (Strong/High grade of evidence).[41] As described in the guidelines, using data from the one randomized controlled trial of treatment extension in 1000 TB patients with a 7% risk of relapse, extending treatment in 183 patients with positive sputum smears at the end of month 2 would prevent 16 of the 70 predicted relapses; however, an additional 158 patients would have had their treatment needlessly extended.[41] The Guideline recommendation was based on the modest benefit of treatment extension, an inability to predict at-risk individuals, and potential downsides to treatment extension that include increased utilization of programmatic resources and the increased potential for medication related side effects.

This systematic review and meta-analysis had several strengths. We used a standard protocol that included a comprehensive search strategy, independent reviewers, assessment of the quality of the included studies, and a hierarchical regression approach for meta-analysis. We presented pooled results using both measures of association (odds ratio) and diagnostic accuracy (sensitivity and specificity). The goal of sputum evaluation during treatment is the prediction of a future event, rather than the identification of an existing condition. However, despite a strong statistical association, a marker may not be able to discriminate between individuals who will or will not experience an outcome. The performance of a marker to predict individual risk is better demonstrated by sensitivity and specificity.[42]

The review also had limitations. Analyses were limited by the small number of included studies for a particular outcome, type of sputum examination, and month of evaluation. There were limited data for HIV co-infected individuals. Drug-susceptibility testing was not performed in all the studies. Several studies enrolled subjects based on a positive sputum culture but in their sputum examinations during treatment reported sputum smear in addition to culture. For the most part, studies were unable to differentiate relapse from exogenous re-infection, which may be an important source of recurrent infection in TB endemic settings.[43]

Studies varied in the number sputum samples that were assessed: 14 (50%) studies did not report the number of specimens examined during treatment. Although statistical tests and graphical methods for the detection of potential publication bias in meta-analyses of randomized control trials are available, to our knowledge such techniques have not been adequately evaluated for diagnostic data.[44] Finally, our search strategy may have missed some relevant studies by excluding non-English publications.

The use of systematic review methods in the area of prognosis is a relatively new area. While basic principles to address bias and random error are similar to those used for other reviews, there are challenges in both identifying studies and combining results across different study designs and analyses. The use of individual patient rather than published data might have improved the quality of our meta-analysis, but this approach requires considerable time and resources, which were unavailable for this project.[45] Of interest, a re-analysis by mixed effects logistic regression of individual patient data from the clinical trials conducted by the British Medical Research Council in the 1970s and 80s has recently been performed. In this analysis, only data from patients on six month, four-drug regimens with isoniazid and rifampin throughout, regimens broadly similar to those in current use, were included. The investigators found sensitivities for culture or smear at two months for predicting relapse to be low (less than 40%), similar to the results in the current review (Patrick PJ Phillips, personal communication 3 October 2008).

Endogenous TB relapses are thought to be due to a failure to eradicate persistent bacilli that are remote from cavitary surfaces (e.g. in necrotic regions of the lung), metabolically less active, and less susceptible to drug therapy.[46] These persisting populations of bacilli may be undetected by standard methods of sputum examination. A recent study of treatment duration (4- versus 6-months) in subjects who had converted their sputum cultures by the end of 2 months of treatment was stopped due to excess relapses in the 4-month arm demonstrating the limitations of sputum culture for predicting patient outcomes.[47]

The poor performance of sputum smear and culture as prognostic markers for poor outcome also raises questions about the use of these markers as surrogate endpoints in clinical trials, in particular phase 2 trials evaluating new TB medications. Although relapse will remain the optimal microbiologic endpoint for phase 3 trials, a reliable surrogate marker of treatment outcome that can be measured early in treatment is needed to fast track phase 2 trials. Surrogate markers that could provide an early indication of drug efficacy and correctly predict regimen specific effects on relapse and/or failure would markedly decrease the cost of clinical trials and accelerate the development of new drugs. To date, two month sputum culture is the most widely studied and used putative surrogate for treatment outcome.[12, 48] However, this meta-analysis found that two month culture was relatively insensitive and had poor PPV for predicting relapse. Whereas this poor performance highlights the need for more sensitive and specific biomarkers of treatment response, it is important to note that alternative, newer biomarkers have yet to be sufficiently validated within phase 3 trials to permit their use as surrogate endpoints of failure and relapse.[12, 49] The development and validation of new surrogate markers could be a secondary aim of clinical trials.[50]

We were surprised by the generally weak evidence base that supports such a fundamental and important component of TB control as sputum monitoring. The widespread use of this test and its uncertain utility suggest a critical need to appraise tests and strategies for identifying patients at risk for poor outcome that are suitable for resource limited settings. As recommended by the GRADE approach, future studies of diagnostic tests and strategies should evaluate outcomes that go beyond sensitivity and specificity by assessing whether these tests and strategies result in improved outcomes that are important to patients.[51]

Although the PPV of a sputum specimen for relapse or failure was low in this review, a positive sputum smear at the end of the intensive phase may be useful for several reasons. First, it should trigger an assessment of the quality of patient support, as well as the dosage and quality of TB medications. Reasons for any treatment interruptions or gaps in adherence should be rapidly explored and addressed. Second, a positive smear at this juncture is now recommended to trigger an additional sputum smear at month 3; if still positive, culture and drug-susceptibility testing are recommended.[41] Finally, the proportion of smear positive patients with sputum smear conversion at the end of the intensive phase is also an indicator of TB program performance. The use of a sputum specimen to predict poor outcome and decide on an extension of treatment, should be questioned until further studies have addressed these issues.

Acknowledgments

The authors would like to thank Megan Henry, Midori Kato-Maeda, Dennis Osmond, Diana Pope and Gloria Won for their support and assistance.

Funding/Support: This project was supported in part by the World Health Organization, Tuberculosis Strategy and Health Systems, Stop TB Department, Geneva. PCH, DJH, and PN receive support through the US National Institutes of Health. MN receives support through the US Centers for Disease Control and Prevention.

References

1. World Health Organization. Global tuberculosis control: epidemiology, planning, financing: WHO report 2009. Geneva: World Health Organization; 2009.
2. World Health Organization. Treatment of Tuberculosis: Guidelines for National Programmes. WHO Global Tuberculosis Programme; 2003.
3. Dye C, Watt CJ, Bleed DM, Hosseini SM, Raviglione MC. Evolution of tuberculosis control and prospects for reducing tuberculosis incidence, prevalence, and deaths globally. JAMA 2005;293:2767–75. [PubMed: 15941807]
4. Sterling TR, Alwood K, Gachuhi R, et al. Relapse rates after short-course (6-month) treatment of tuberculosis in HIV-infected and uninfected persons. AIDS 1999;13:1899–904. [PubMed: 10513648]
5. Hong Kong Chest Service/British Medical Research Council. Five-year follow-up of a controlled trial of five 6-month regimens of chemotherapy for pulmonary tuberculosis. Am Rev Respir Dis 1987;136:1339–42. [PubMed: 2891333]
6. Wilkinson D, Bechan S, Connolly C, Standing E, Short GM. Should we take a history of prior treatment, and check sputum status at 2–3 months when treating patients for tuberculosis? Int J Tuberc Lung Dis 1998;2:52–5. [PubMed: 9562111]
7. Enarson DA, Jindani A, Kuaban C, et al. Appropriateness of extending the intensive phase of treatment based on smear results. Int J Tuberc Lung Dis 2004;8:114–6. [PubMed: 14974754]
8. Benator D, Bhattacharya M, Bozeman L, et al. 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. [PubMed: 12241657]
9. Jindani A, Nunn AJ, Enarson DA. Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. Lancet 2004;364:1244–51. [PubMed: 15464185]
10. Mitchison DA. Shortening the treatment of tuberculosis. Nat Biotechnol 2005;23:187–8. [PubMed: 15696148]
11. Zhao FZ, Levy MH, Wen S. Sputum microscopy results at two and three months predict outcome of tuberculosis treatment. Int J Tuberc Lung Dis 1997;1:570–2. [PubMed: 9487456]
12. Wallis RS, Doherty TM, Onyebujoh P, et al. Biomarkers for tuberculosis disease activity, cure, and relapse. Lancet Infect Dis 2009;9:162–72. [PubMed: 19246020]
13. World Health Organization. Treatment of Tuberculosis Guidelines. 2009. http://www.who.int/tb/publications/2009/who_htm_tb_2009_420_beforeprint.pdf

14. Deville WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2:9. [PubMed: 12097142]
15. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol 2006;187:271–81. [PubMed: 16861527]
16. 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:A11–3. [PubMed: 15230574]
17. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889–97. [PubMed: 19075208]
18. Rom, WN.; Garay, SM. Tuberculosis. 2. Philadelphia: Lippincott Williams & Wilkins; 2004.
19. Shingadia D, Novelli V. Diagnosis and treatment of tuberculosis in children. Lancet Infect Dis 2003;3:624–32. [PubMed: 14522261]
20. Cao JP, Zhang LY, Zhu JQ, Chin DP. Two-year follow-up of directly-observed intermittent regimens for smear-positive pulmonary tuberculosis in China. Int J Tuberc Lung Dis 1998;2:360–4. [PubMed: 9613630]
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. [PubMed: 14606960]
22. 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. [PubMed: 18436948]
23. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–84. [PubMed: 11568945]
24. 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. [PubMed: 18253484]
25. Harbord, R.; Whiting, P. Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. In: Sterne, J., editor. Meta-analysis in Stata. StataCorp LP; 2009.
26. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006;6:31. [PubMed: 16836745]
27. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002;21:1525–37. [PubMed: 12111918]
28. Chang KC, Leung CC, Yew WW, Ho SC, Tam CM. A nested case-control study on treatment-related risk factors for early relapse of tuberculosis. Am J Respir Crit Care Med 2004;170:1124–30. [PubMed: 15374844]
29. Nettles RE, Mazo D, Alwood K, et al. Risk factors for relapse and acquired rifamycin resistance after directly observed tuberculosis treatment: a comparison by HIV serostatus and rifamycin use. Clin Infect Dis 2004;38:731–6. [PubMed: 14986259]
30. Picon PD, Bassanesi SL, Caramori ML, Ferreira RL, Jarczewski CA, Vieira PR. Risk factors for recurrence of tuberculosis. J Bras Pneumol 2007;33:572–8. [PubMed: 18026656]
31. Tam CM, Chan SL, Kam KM, Goodall RL, Mitchison DA. Rifapentine and isoniazid in the continuation phase of a 6-month regimen. Final report at 5 years: prognostic value of various measures. Int J Tuberc Lung Dis 2002;6:3–10. [PubMed: 11931398]
32. Thomas A, Gopi PG, Santha T, et al. Predictors of relapse among pulmonary tuberculosis patients treated in a DOTS programme in South India. Int J Tuberc Lung Dis 2005;9:556–61. [PubMed: 15875929]
33. Ramarokoto H, Randriamiharisoa H, Rakotoarisaonina A, et al. Bacteriological follow-up of tuberculosis treatment: a comparative study of smear microscopy and culture results at the second month of treatment. Int J Tuberc Lung Dis 2002;6:909–12. [PubMed: 12365578]
34. Van Deun A, Aung KJ, Hamid Salim MA, et al. Extension of the intensive phase reduces unfavourable outcomes with the 8-month thioacetazone regimen. Int J Tuberc Lung Dis 2006;10:1255–61. [PubMed: 17131785]
35. Dembele SM, Ouedraogo HZ, Combary A, Saleri N, Macq J, Dujardin B. Conversion rate at two-month follow-up of smear-positive tuberculosis patients in Burkina Faso. Int J Tuberc Lung Dis 2007;11:1339–44. [PubMed: 18034956]

36. Rieder HL. Sputum smear conversion during directly observed treatment for tuberculosis. Tuber Lung Dis 1996;77:124–9. [PubMed: 8762846]
37. Santha T, Garg R, Frieden TR, et al. Risk factors associated with default, failure and death among tuberculosis patients treated in a DOTS programme in Tiruvallur District, South India, 2000. Int J Tuberc Lung Dis 2002;6:780–8. [PubMed: 12234133]
38. Zierski M, Bek E, Long MW, Snider DE Jr. Short-course (6 month) cooperative tuberculosis study in Poland: results 18 months after completion of treatment. Am Rev Respir Dis 1980;122:879–89. [PubMed: 7006476]
39. Koepsell, TD.; Weiss, NS. Epidemiologic methods: studying the occurrence of illness. Oxford; New York: Oxford University Press; 2003.
40. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials 2006;3:272–9. [PubMed: 16895044]
41. World Health Organization. Treatment of Tuberculosis Guidelines. 2009. http://www.who.int/tb/publications/2009/who_htm_tb_2009_420_beforeprint.pdf
42. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–90. [PubMed: 15105181]
43. Verver S, Warren RM, Beyers N, et al. Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. Am J Respir Crit Care Med 2005;171:1430–5. [PubMed: 15831840]
44. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005;142:1048–55. [PubMed: 15968029]
45. Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. Br J Cancer 2009;100:1219–29. [PubMed: 19367280]
46. Boshoff HI, Barry CE 3rd. Tuberculosis - metabolism and respiration in the absence of growth. Nat Rev Microbiol 2005;3:70–80. [PubMed: 15608701]
47. Johnson JL, Hadad DJ, Dietze R, et al. Shortening treatment in adults with noncavitary tuberculosis and 2-month culture conversion. Am J Respir Crit Care Med 2009;180:558–63. [PubMed: 19542476]
48. Mitchison DA. Assessment of new sterilizing drugs for treating pulmonary tuberculosis by culture at 2 months. Am Rev Respir Dis 1993;147:1062–3. [PubMed: 8466107]
49. Perrin FM, Lipman MC, McHugh TD, Gillespie SH. Biomarkers of treatment response in clinical trials of novel antituberculosis agents. Lancet Infect Dis 2007;7:481–90. [PubMed: 17524807]
50. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008;100:1432–8. [PubMed: 18840817]
51. 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. [PubMed: 18483053]

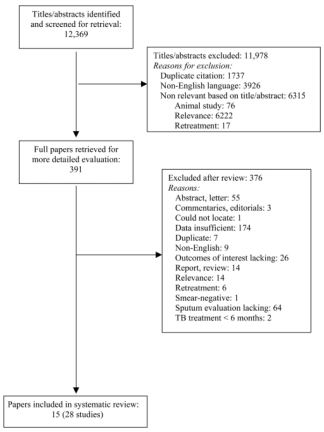

Figure 1.
Study selection process. TB: tuberculosis.

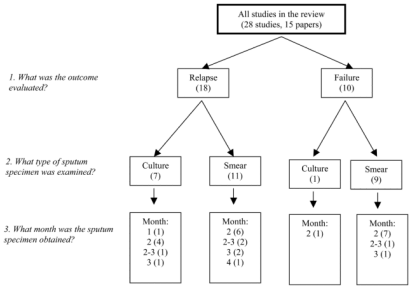

Figure 2.
Sub-group Selection.

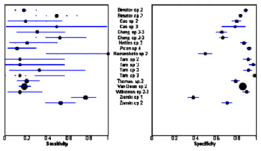

Figure 3.
Positive sputum specimen as a predictor of relapse. The circles and the lines represent the point estimates and 95% CIs, respectively. The size of the circle indicates the study size. Cp, culture positive; sp, sputum smear positive; the number following cp or sp indicates the month the sputum specimen was examined. Sensitivity is the proportion of relapsed subjects who had a positive sputum examination. Specificity is the proportion of subjects who did not relapse and had a negative sputum examination.

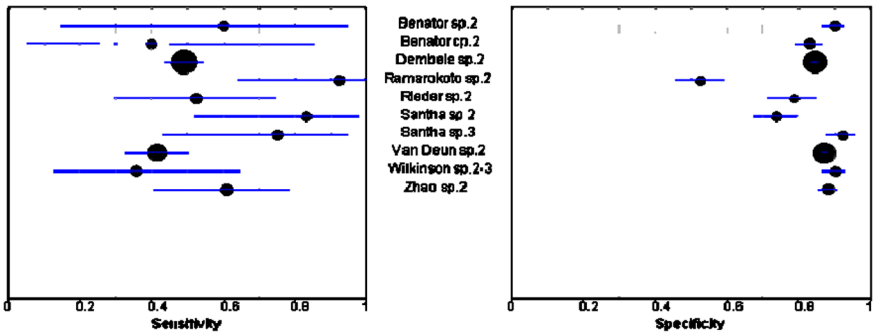

Figure 4.

Positive sputum specimen as a predictor of failure. The circles and the lines represent the point estimates and 95% CIs, respectively. The size of the circle indicates the study size. Cp, culture positive; sp, sputum positive; the number following cp or sp indicates the month the sputum specimen was examined. Sensitivity is the proportion of subjects who experienced treatment failure and had a positive sputum examination. Specificity is the proportion of subjects who did not experience treatment failure and had a negative sputum examination.

Table 1

Assessment of Study Quality

Characteristic	Number
Study design	
Cohort	23
Case-control	5
Patient selection	
Consecutive	13
Random	12
Convenience or not reported	3
Were fewer than 10% of patients lost during treatment? [1]	
Yes	11
No	11
Not applicable/unclear[2]	6
Were fewer than 10% of patients lost during follow-up?[3]	
Yes	11
No	1
Not applicable/unclear[4]	6
Manner of follow-up for relapse (18 studies) [5]	
Active	12
Passive	3
Not applicable[6]	3
Supervision of treatment fully observed	
Yes	22
No	5
Not reported	1
Outcome culture confirmed	
Yes	11
No	17

[1] Defined as time from treatment initiation to completion.

[2] Includes five case-control studies and one study that did not report the number of patients enrolled.

[3] Defined as 12 months after treatment completion.

[4] Includes three case-control studies and three studies with passive follow-up of patients.

[5] Patients were followed for at least 12 months after treatment completion.

[6] Quality criterion considered not applicable for three case-control studies.

Table 2a

Characteristics of included studies, sputum specimen predicting relapse

First author (area) [reference]	Study design	Patient selection	Reference for TB diagnosis	Treatment regimen	Specimen type	Month specimen examined	Number specimen+/Total	Sensitivity (95% CI)	Specificity (95% CI)
Benator (USA & Canada) [8]	Cohort	Random	Culture	$2HRZE4(RH)_2$	Smear	2	44/413	18 (5–40)	90 (86–93)
Benator (USA & Canada) [8]	Cohort	Random	Culture	$2HRZE4(RH)_2$	Culture	2	72/418	50 (28–72)	85 (81–88)
Cao (China) [17]	Cohort	Consecutive	Smear	$2(HRZE)_34(HR)_3$ $2(HRZE)_36(HRE)_3$	Smear	2	56/296	20 (3–56)	81 (76–85)
Cao (China) [17]	Cohort	Consecutive	Smear	$2(HRZE)_34(HR)_3$ $2(HRZE)_36(HRE)_3$	Smear	3	12/56	50 (1–99)	80 (66–89)
Chang (Hong Kong) [26]	Case-control	Not reported	Bacteriologic and/or clinical	2HRZ4HR (daily or 3/week†)	Smear	2–3	113/339	31 (11–59)	67 (61–72)
Chang (Hong Kong) [26]	Case-control	Not reported	Bacteriologic and/or clinical	2HRZ4HR (daily or 3/week†)	Culture	2–3	113/339	53 (27–79)	68 (62–73)
Nettles (USA) [27]	Cohort	Consecutive	Culture	$2(HRZE)_2(HR)_2$*	Culture	2	45/361	21 (5–51)	88 (84–91)
Picon (Brazil) [28]	Cohort	Consecutive	Smear	2HRZ4HR 2HRZ7HR	Smear	4 or later	43/610	12 (2–30)	93 (91–95)
Ramarokoto (Madagascar) [31]	Case-control	Consecutive	Culture or smear	2HRZE6HE**	Smear	2	117/234	100 (40–100)	51 (44–57)
Tam (Hong Kong) [29]	Cohort	Random	Culture	$2SHRZ4(HR)_3$	Smear	2	14/172	14 (0–58)	92 (87–96)
Tam (Hong Kong) [29]	Cohort	Random	Culture	$2SHRZ4(HR)_3$	Smear	3	6/172	14 (0–58)	97 (93–99)
Tam (Hong Kong) [29]	Cohort	Random	Culture	$2SHRZ4(HR)_3$	Culture	2	14/167	33 (4–78)	93 (87–96)
Tam (Hong Kong) [29]	Cohort	Random	Culture	$2SHRZ4(HR)_3$	Culture	3	4/167	14 (0–58)	98 (95–100)
Thomas (India) [30]	Cohort	Consecutive	Smear	$2(HRZE)_34(HR)_3$	Smear	2	100/503	21 (12–33)	80 (76–84)
Van Deun (Bangladesh) [18]	Cohort	Random	Culture	2HRZE6HT	Smear	2	1098/8230	19 (14–25)	87 (86–88)
Wilkinson (South Africa) [6]	Cohort	Consecutive	Culture	$6(HRZS)_2$	Smear	2 or 3	31/320	14 (3–36)	91 (87–94)
Zierski (Poland) [35]	Cohort	Random	Culture	6HRE daily $2HRE4(HR)_2$ $2HRE4(HRE)_1$ $2HRE4(HRE)_2$	Culture	1	222/352	78 (64–88)	39 (34–45)
Zierski (Poland) [35]	Cohort	Random	Culture	6HRE daily $2HRE4(HR)_2$ $2HRE4(HRE)_1$ $2HRE4(HRE)_2$	Culture	2	112/352	54 (39–68)	72 (66–77)

H- isoniazid, R- rifampin, Z- pyrazinamide, E- ethambutol, S- streptomycin, T- thiacetazone

† Chang regimens: Various other regimens included rifampin-containing regimens with subsequent omission of isoniazid and extended use of ethambutol and pyrazinamide; 9-month regimen (containing mainly isoniazid and rifampicin); regimens with rifapentine during continuation phase, and rifamycin-deficient regimens.

* Duration of continuation phase was determined by clinician

** Treatment extended if smear-positive at 2 months

Table 2b

Characteristics of included studies, sputum specimen predicting failure

First author (area) [reference]	Study design	Patient selection	Reference for TB diagnosis	Treatment regimen	Specimen type	Month specimen examined	Number specimen+/Total	Sensitivity (95% CI)	Specificity (95% CI)
Benator (USA & Canada) [8]	Cohort	Random	Culture	$2HRZE4(RH)_2$	Smear	2	44/413	60 (15–95)	90 (87–93)
Benator (USA & Canada) [8]	Cohort	Random	Culture	$2HRZE4(RH)_2$	Culture	2	72/418	40 (5–85)	83 (79–87)
Dembele (Burkina Faso) [32]	Cohort	Consecutive	Smear	2HRZE6HE	Smear	2	1688/10,054	49 (43–54)	84 (84–85)
Ramarokoto (Madagascar) [31]	Case- control	Consecutive	Culture or smear	2HRZE6HE*	Smear	2	117/234	92 (64–100)	52 (46–59)
Rieder (Thailand) [33]	Case- control	Consecutive	Smear	2HRSZ4HR	Smear	2	44/176	52 (30–74)	79 (71–85)
Santha (India) [34]	Cohort	Consecutive	Smear	$2(HRZE/HR)_3$	Smear	2	67/229	83 (52–98)	74 (67–79)
Santha (India) [34]	Cohort	Consecutive	Smear	$2(HRZE/HR)_3$	Smear	3	26/229	75 (43–95)	92 (88–95)
Van Deun (Bangladesh) [18]	Cohort	Random	Culture	2HRZE6HT	Smear	2	1098/8230	41 (33–50)	87 (86–88)
Wilkinson (South Africa) [6]	Cohort	Consecutive	Culture	$6(HRZS)_2$	Smear	2 or 3	40/365	36 (13–65)	90 (86–93)
Zhao (China) [11]	Cohort	Not reported	Smear	$2(HRZS)_34(HR)_3$	Smear	2	100/726	61 (41–78)	88 (85–90)

H- isoniazid, R- rifampin, Z- pyrazinamide, E- ethambutol, S- streptomycin, T - thiacetazone

*Treatment extended if smear-positive at 2 months

Table 3

Pooled summary estimates for relapse or failure for patients with a positive sputum culture or smear at 2 months

Subgroup	Sample Size (studies)	Hierarchical Regression Model		Odds Ratio	PPV* (95% CI)	NPV* (95% CI)
		Sensitivity (95% CI)	Specificity (95% CI)			
Relapse:						
Culture	1298 (4)	40 (25–56)	85 (77–91)	3.8 (2.2–6.8)	18 (14–21)	95 (95–96)
Smear	9848 (6)	24 (12–42)	83 (72–90)	1.5 (1.1–2.2)	10 (8–12)	93 (93–94)
Failure:						
Smear	20062 (7)	57 (41–73)	81 (72–87)	5.8 (4.3–7.8)	9 (9–10)	98 (98–98)

*Ability of smear to predict poor outcomes assuming 7% risk of relapse and 3% risk of failure.

Appendix

PubMed search strategy run May 21, 2008

Search	Most Recent Queries
#16	Search **#13 NOT #14**
#15	Search **#14 NOT (MURINE[TI] OR MOUSE[TI] OR MICE[TI])**
#14	Search **#5 OR #7 OR #10 OR #12** Limits: **English**
#13	Search **#5 OR #7 OR #10 OR #12**
#12	Search **#11 AND (#4 OR #6)**
#11	Search **PULMONARY TUBERCULOSIS/DRUG THERAPY[MAJR] OR (PULMONARY TUBERCULOSIS[TIAB] AND (TREATMENT OR THERAPY)) AND ((SMEAR[TIAB] OR SMEARS[TIAB]) AND (CULTURE[TIAB] OR CULTURES[TIAB]))**
#10	Search **#1 AND #8 AND #9**
#9	Search **((CLINICAL[TIAB] AND (TRIAL[TIAB] OR TRIALS[TIAB])) OR CLINICAL TRIALS[MH] OR CLINICAL TRIAL[PT] OR RANDOM*[TIAB] OR RANDOM ALLOCATION[MH] OR EXPERIMENTAL STUD*[TIAB])**
#8	Search **RIFAMPIN OR RIFAMPICIN OR RIFAMYCIN OR RIFAMYCINS OR RIFATER OR RIFAMATE OR ISONIAZID OR PYRAZINAMIDE OR ETHAMBUTOL OR ANTITUBERCULAR AGENTS[MAJR:NOEXP]**
#7	Search **#3 AND #6**
#6	Search **(CLINICAL[TIAB] AND (TRIAL[TIAB] OR TRIALS[TIAB])) OR CLINICAL TRIALS[MH] OR CLINICAL TRIAL[PT] OR RANDOM*[TIAB] OR RANDOM ALLOCATION[MH] OR EXPERIMENTAL STUD*[TIAB] OR THERAPEUTIC USE[SH]**
#5	Search **#3 AND #4**
#4	Search **INCIDENCE[MH:NOEXP] OR MORTALITY[MH] OR FOLLOW UP STUDIES[MH:NOEXP] OR PROGNOS* OR PREDICT* OR COURSE***
#3	Search **#1 AND #2**
#2	Search **SPUTUM/MICROBIOLOGY OR SPUTUM/CYTOLOGY OR RECURRENCE[MH] OR TREATMENT FAILURE[MH] OR RELAPSE[TIAB]**
#1	Search **TUBERCULOSIS[MH] OR MYCOBACTERIUM TUBERCULOSIS[MH] OR TUBERCULOSIS[TI]**