Published in final edited form as:

Acad Radiol. 2010 February ; 17(2): 157–162. doi:10.1016/j.acra.2009.08.013.

Effect of Improving the Quality of Radiographic Interpretation on the Ability to Predict Pulmonary Tuberculosis Relapse

Jason E. Stout, MD, MHS, Andrzej S. Kosinski, PhD, Carol Dukes Hamilton, MD, MHS, Philip C. Goodman, MD, Ann Mosher, MPH, FNP, Dick Menzies, MD, Neil Schluger, MD, Awal Khan, PhD, John L. Johnson, MD, and the Tuberculosis Trials Consortium

Division of Infectious Diseases & International Health, Department of Medicine (J.E.S., C.D.H., A.M.), Department of Biostatistics and Bioinformatics (A.S.K.), and Division of Thoracic Imaging, Department of Radiology (P.C.G.), 3306-Duke University Medical Center, Durham, NC 27710; Montreal Chest Institute, McGill University, Montreal, Quebec, Canada (D.M.); Division of Pulmonary Allergy & Critical Care Medicine, Columbia University College of Physicians & Surgeons, New York, NY (N.S.); Centers for Disease Control and Prevention, Atlanta, GA (A.K.); Division of Infectious Diseases, Case Western Reserve University, Cleveland, OH (J.L.J.); the Tuberculosis Trials Consortium

Abstract

Rationale and Objectives—Chest radiographic findings are important for diagnosis and management of tuberculosis. The reliability of these findings is therefore of interest. We sought to describe interobserver reliability of chest radiographic findings in pulmonary tuberculosis, and to understand how the reliability of these findings might affect the utility of radiographic findings in predicting tuberculosis relapse.

Materials and Methods—Three blinded readers independently reviewed chest radiographs from a randomly selected group of 10% of HIV-seronegative subjects participating in a tuberculosis treatment trial. The three readers then arrived at a fourth, consensus radiographic interpretation.

Results—A total of 241 films obtained from 99 patients were reviewed. Agreement among the independent readers was very good for the findings of bilateral disease ($\kappa = 0.71$ – 0.86 among readers) and cavitation ($\kappa = 0.66$ – 0.73). The original interpretation was reasonably sensitive and specific (compared to the consensus interpretation) for bilateral disease, but the sensitivity for cavity decreased from 81% for the 2-month film to 47% at end of treatment ($P = 0.013$). Substituting the consensus interpretation for the original interpretation increased the odds ratio for the association between cavitation on early chest radiograph and subsequent tuberculosis relapse from 4.97 to 8.97.

Conclusion—Radiographic findings were reasonably reliable between independent reviewers and the original interpretations. The original investigators, who knew the patient's clinical course, were less likely to identify cavitation on the end of treatment chest radiograph. Improving the reliability of these findings could improve the utility of chest radiographs for predicting tuberculosis relapse.

Keywords

Tuberculosis; radiography; thoracic; reliability and validity

Chest radiographic findings have traditionally been important for diagnosis and management of tuberculosis (TB). The current American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America guidelines for TB treatment recommend extending treatment from 6 to 9 months for patients with both a cavity on early (taken after 2 months or less of treatment) chest radiograph (CXR) and a positive sputum culture after 2 months of antituberculous therapy (1). This recommendation was based to a large extent on the findings of the Tuberculosis Trials Consortium Study 22. Study 22 was a multicenter, randomized, controlled trial that compared a standard twice-weekly TB treatment regimen during the continuation phase of therapy with a once-weekly regimen substituting a new drug, rifapentine, for the standard agent, rifampin. The final study cohort of 1004 subjects was restricted to HIV-seronegative individuals (13). Five characteristics were found to be independently associated with increased risk for disease relapse: positive sputum cultures for *Mycobacterium tuberculosis* after 2 months of antituberculous therapy; the presence of a cavity on a CXR taken during the first 2 months of treatment, being underweight, having bilateral pulmonary involvement, and being a non-Hispanic white person. The risk for TB relapse was 23% in the group with both culture-positive sputum after 2 months of therapy plus a cavity on an early CXR, and less than 2% in the group with neither a cavity on chest film nor positive sputum cultures after 2 months of antituberculous therapy (2).

There are several reasons why it is important to better understand the reliability of CXR interpretations. First, there are now clinical ramifications for recognizing or missing the presence of a cavity on an early (taken either at initiation of treatment or at 2 months after initiation of treatment) CXR. Second, in response to the findings of Study 22, clinical trials examining new therapeutic options for treatment of TB are basing enrollment strategies and analyses on subjects' radiographic pattern. Unfortunately, several studies have described only moderate agreement among independent observers in evaluating radiographs for manifestations of TB (3–7). Furthermore, agreement may vary depending on the observer's specialty and level of experience (3,8). When the individual interpreting the chest radiograph is familiar with the patient's clinical condition, the interpretation of the chest radiograph may also be biased by the clinical data. Bias and interreader variability may both affect the utility of radiographic data in predicting TB outcomes. We therefore conducted a substudy to validate the radiographic findings of Study 22 and to assess how interreader reliability and bias might have affected the association between radiographic findings and TB relapse.

MATERIALS AND METHODS

We conducted a validation substudy of United States Public Health Service Study 22, a randomized, controlled trial comparing standard twice-weekly isoniazid and rifampin to once-weekly isoniazid and rifapentine during the continuation phase of TB treatment in HIV-seronegative subjects (2). Data from up to three standard posteroanterior chest radiographs were obtained from each subject. Chest radiographs were required within 2 weeks of randomization, which occurred after 8 weeks of standard therapy, and at the end of TB treatment. A CXR at the time of initial diagnosis was not required for the study, but if available, this radiograph was assessed in the same way as the study-required films and included in the study database. Central reading of CXRs was not performed during Study 22. The reported findings were based only on standard posteroanterior and lateral chest radiographs. Although site investigators were either chest or infectious disease physicians with considerable experience in the diagnosis and management of TB, site personnel did not receive specific study-related training on CXR interpretation or study reporting before or during Study 22. Specifically, site investigators were asked to report presence or absence of cavity based on their judgment and experience, as well as interpretations from local

radiologists, but the term “cavity” was not otherwise defined. For each chest radiograph, site investigators reported three binary variables: 1) normal/abnormal; 2) presence/absence of bilateral TB disease; and 3) presence/absence of one or more pulmonary cavities. For the second and third radiographs, site investigators also reported whether there was extension of TB lesions present on a previous radiograph, and whether new TB lesions were present compared to a previous radiograph.

A simple, random sample of 20% of the original 1004 Study 22 subjects was selected for the validation substudy. The original Study 22 radiographs were mailed to Duke University Medical Center, where subject, radiograph date, and site identifiers were obscured by study personnel not involved in interpreting the radiographs. Three readers (one chest radiologist, PG, and two infectious diseases/TB clinicians, CDH and JES) independently interpreted the radiographs using the original five Study 22 variables (normal, bilateral disease, cavity, extension of old lesions, and new lesions). Readers were presented the films in random order and blinded to the original order of the films as well as to all clinical and microbiological data. All radiographs for each subject (up to three) were read at the same time by each reader, and readers were asked to make pairwise comparisons of all films to assess extension of old lesions and appearance of new lesions. After all three readers had independently interpreted the radiographs, a consensus meeting was held to resolve any differences of opinion regarding radiograph interpretation. The consensus interpretation, which was held by at least a majority of readers, was held to be the “gold standard” for purposes of this analysis.

Statistical Methods

We used the kappa statistic to describe agreement between radiographic interpretations and McNemar’s chi-square test to test whether systematic differences (biases) were present between the blinded and original site radiographic interpretations (9). We compared the original sites’ interpretations to the blinded consensus interpretations and summarized the findings by sensitivity and specificity, using the blinded consensus readings as the referent or gold standard. We then used a latent class model, with an assumption of independence of the blinded readings conditional on TB relapse status, to assess the validity of the consensus reading in the sampled chest radiographs (10,11). The relationship between cavity status (present/absent) and TB relapse status was summarized by odds ratios (OR) for the original site readings and for available blinded consensus readings. To estimate the relationship between the “gold standard” presence of cavity and relapse we also utilized a maximum likelihood approach (12), which simultaneously provides OR estimates for associations between TB relapse and the true cavity status (as determined by blinded consensus), as well as estimates of sensitivity and specificity of the original site readings with the consensus cavity status as the gold standard. This approach uses in one model all the original Study 22 readings and available consensus readings for those in the validation substudy. We considered a model with nondifferential misclassification of true cavity status by the original site reading, and with the missing at random structure (13), which states that the availability of the blinded consensus readings is not related to the TB relapse status. Using this model, we created curves that estimated the OR associating cavity on early chest radiograph with relapse as a function of sensitivity and specificity of cavity reading (as compared to the consensus gold standard) from the formula:

$$OR(sensitivity, specificity) = \frac{\theta_{Relapse} / (1 - \theta_{Relapse})}{\theta_{No\ relapse} / (1 - \theta_{No\ relapse})},$$

where

$$\theta_{\text{Relapse}} = \theta_{\text{Relapse}}^{\text{True}} \times \text{sensitivity} + (1 - \theta_{\text{Relapse}}^{\text{True}}) \times (1 - \text{specificity})$$

and

$$\theta_{\text{No relapse}} = \theta_{\text{No relapse}}^{\text{True}} \times \text{sensitivity} + (1 - \theta_{\text{No relapse}}^{\text{True}}) \times (1 - \text{specificity})$$

Relapse represents the proportion of subjects with a cavity, as determined by an observer with sensitivity/specificity as noted, who relapse after TB treatment. $\theta_{\text{Relapse}}^{\text{True}}$ represents the proportion of subjects who “truly” have a cavity, as determined by a theoretical perfect observer, and relapse after TB treatment. No relapse represents the proportion of subjects with a cavity, as determined by an observer with sensitivity/specificity as noted, who do not relapse after TB treatment. $\theta_{\text{No relapse}}^{\text{True}}$ represents the proportion of subjects who “truly” have a cavity, as determined by a theoretical perfect observer, and do not relapse after TB treatment. Sensitivity was defined as the proportion of radiographs for which a given observer detected a cavity when a cavity was “truly” present. Specificity was defined as the proportion of radiographs for which a given observer did not detect a cavity when a cavity was not “truly” present.

Statistical significance for all analyses was defined as a two-sided *P* value of <0.05. Data analysis was performed with SAS 9.1 (SAS Systems, Cary, NC) and Splus 6.1.2 (Insightful Corp, Seattle, WA). The study was approved by the institutional review board at the Centers for Disease Control and Prevention as well as the participating institutions.

RESULTS

Two hundred and two of the original Study 22 subjects were randomly selected to participate in the validation substudy. Of these, at least one radiograph could be obtained for 99 of the 202 (9.9% of the total Study 22 population). Radiographs were not available for all three time points (start of treatment, 2-month, end of treatment) for every subject, so a total of 241 radiographs (of a possible 297) were obtained and read by each of the three blinded readers. Demographic characteristics of the 99 subjects for whom chest radiographs could be obtained were not significantly different than the rest of the Study 22 population, and represented both study arms (45 were randomized to rifapentine, 54 to rifampin).

Interreader agreement among the three blinded readers for the entire set of 241 radiographs is described in Table 1. The readers generally agreed on whether a given film was normal (93%–95% agreement between pairs of readers), bilateral (85%–93%), or had a cavity (84%–88%). Kappa values, representing agreement over what would be expected by chance, were low for the assessment of normal/abnormal film (0.35–0.45) because the vast majority of films were read as abnormal. By random chance one would expect a high proportion of agreement for this variable, so a low absolute level of disagreement results in a relatively low kappa value for this variable. Kappa values suggested substantial to near-perfect agreement for the finding of bilateral disease ($\kappa = 0.71$ –0.86) and substantial agreement for the finding of cavity ($\kappa = 0.66$ –0.73) (14). When comparing pairs of films for the same subject, agreement on whether there was extension of old lesions ranged from 85%–87% ($\kappa = 0.70$ –0.75), whereas agreement on the presence of a new lesion on one CXR that was not present on another ranged between 82% and 90%; kappa values were low for this measure ($\kappa = 0.28$ –0.47) because new lesions were a very uncommon finding. The blinded readers

interpreted chest radiographs over 235 days. The magnitude of interreader agreement for normal/abnormal, bilateral disease, and cavity did not significantly change between the first half of this period and the last half.

We used a latent class model, which does not require a gold standard measurement, to assess the validity of using a consensus session to arrive at a “gold standard” interpretation, and estimated the sensitivity and specificity of each reader for the finding of cavity on the 2-month CXRs. The sensitivity and specificity of each reader as estimated by a direct comparison to the consensus interpretation and as estimated by the latent class model are described in Table 2. The sensitivity and specificity of each reader as compared to the consensus reading were within 5% of the sensitivity and specificity estimated by the latent class model for all readers.

Given the good performance of the consensus reading, we examined agreement between the original radiograph interpretation at the study site and the consensus interpretation, being used now as the “gold standard” (Table 3). For all CXRs, the original site interpretations were 97% sensitive and 71% specific for the determination of any abnormality on the CXRs, 81% sensitive and 94% specific for bilateral disease, and 65% sensitive and 88% specific for the presence of cavity. The sensitivity of the original site’s interpretation for cavity varied depending on the timing of the radiograph; sensitivity for cavity was significantly higher after 2 months of treatment than at end of treatment (81% vs. 47%, $P < .05$). Sensitivity of the original interpretation for abnormality and bilateral disease did not significantly vary with timing of the radiograph, nor did specificity of the original interpretations for abnormality, bilateral disease, or cavity (all $P > .05$). When discrepancies were present between the original and consensus interpretations, the original interpretation was less likely to call bilateral disease than the consensus interpretation (ie, biased) (McNemar’s test $P = .0004$). The original interpretation was not biased with regard to the finding of cavity on the initial and 2-month radiographs, but was significantly less likely to call a cavity on the end of treatment radiograph than the consensus interpretation ($P = .016$).

In the original Study 22 dataset of 1004 subjects, the finding of cavity on early CXR (either at initiation of treatment or at 2 months after initiation of treatment) was highly associated with subsequent TB relapse. In univariate analysis, persons with a cavity on early (either the initial or 2 month) chest radiograph had an OR of 4.97 (95% CI 2.5–9.9) to relapse compared with persons without a cavity on early CXR. The OR of 4.97 is an estimate of the association between the relapse and the finding of cavity from the original Study 22 radiographic interpretation. Assuming that the consensus reading is a better assessment of the presence or absence of cavity than the original Study 22 reading, we performed two additional analyses to approximate the “true” association between cavity on early chest radiograph and TB relapse. First, we considered only the 99 patients in the validation substudy for whom we had consensus readings and examined the direct relationship between the finding of cavity on the consensus interpretation and relapse. We found an OR of 8.97 (95% CI 1.0–80.5) for relapse among persons with a cavity on early chest radiograph using this approach. Second, we used all the original Study 22 subjects with available early cavity readings ($n = 976$) in the maximum likelihood model approach (12) and found an OR of 12.42 (95% CI 2.9–53.6) for relapse with a cavity on consensus early chest radiograph using this second estimation.

In practice, it is unlikely that consensus interpretations could be used for all radiographs, but it is possible that reader performance could be improved with training. We examined the potential effect of such training. We assumed that training may increase the sensitivity of the readers (ie, readers will detect a larger proportion of cavities that are truly present), the specificity of the readers (ie, readers will make fewer false positive calls of cavities when no

cavity is in fact present), or both. The baseline assumptions used for this analysis, derived from the maximum likelihood analysis, were that the true prevalence of cavitation among subjects with TB relapse is 89.34% ($\theta_{\text{Relapse}}^{\text{True}} = 0.8934$), and that the true prevalence of cavitation among subjects without TB relapse is 40.3% ($\theta_{\text{No relapse}}^{\text{True}} = 0.4030$). Figure 1 shows the effect of improving the sensitivity and specificity of a hypothetical reader (as compared to the “gold standard” consensus reading) on the association (OR) between cavity on early chest radiograph and TB relapse. The maximum likelihood model estimates of the sensitivity and specificity of the original cavity readings (as compared to the consensus) are 89.5% and 73.8%, respectively. The OR for the association between cavity and relapse given the maximum likelihood model estimates of sensitivity and specificity is 4.42 (triangle in Figure 1), which is quite similar to the OR of 4.97 actually observed from direct tabulation of the original Study 22 data. The similarity of these results supports the validity of the maximum likelihood model. Improvement in sensitivity or specificity of the finding of cavity results in a higher predicted OR for relapse given a cavity on early CXR. Taking this improvement to the limit (ie, sensitivity and specificity of 100%) suggests that the maximum possible OR for relapse given a cavity that could be obtained by a perfect radiologist is 12.42 (square in Figure 1).

DISCUSSION

The chest radiographic manifestations of pulmonary TB are useful both diagnostically and prognostically, so being able to accurately recognize and identify these manifestations is of great interest to both clinicians and public health programs in areas where chest radiography is used as an adjunct to TB diagnosis. The accuracy of the finding of cavity on chest radiographs is particularly important, because finding a cavity is associated with an approximately threefold higher risk of TB relapse. Because of the increased risk of relapse, the American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America guidelines currently recommend that persons with a cavity and a positive sputum culture after 2 months of standard TB treatment receive extended TB therapy for 9 months (as opposed to 6 months for persons without these two characteristics) (1). We found that agreement regarding this finding among three blinded observers was substantial (kappa, 0.66–0.73), and that consensus interpretation served as a reasonable gold standard. As compared to this gold standard, unblinded interpretations of the same radiographs across 29 sites by a heterogeneous group of readers were moderately sensitive (70% for initial radiograph) and specific (86% for initial radiograph) for the finding of cavity. However, the sensitivity of these unblinded interpretations on the end-of-treatment CXR fell to 46%. This decrease in sensitivity for cavity on the end-of-treatment CXR is concerning because the finding of cavity on end-of-treatment CXRs has been associated with a significantly increased risk of TB relapse in some (15–18, unpublished data) but not all studies (19). A decrease in sensitivity may be attributable to reduction in cavity size and prominence with treatment, which would make a cavity more difficult to see. Alternatively, this decreased sensitivity may be attributable to bias on the part of the treating clinician, who may have expected the cavity to resolve as part of successful treatment.

Our findings are consistent with other observations on the interobserver reliability of chest radiographs. A large study of 1099 radiographs selected from the 1962 mass screening of the adult population of Oslo, Norway, examined interobserver agreement among 90 readers for key findings associated with TB (20). Agreement was moderate for the findings of any abnormality in the lung and cavity (3). A more recent study based in Montreal examined the performance of three readers in chest radiograph examination in a heterogeneous population of persons being screened for TB (6). They found interobserver agreement, as measured by the kappa statistic, of 0.561 for the finding of abnormal CXR, and 0.491 for the finding of

parenchymal disease with or without cavities. Intra-observer agreement (agreement by the same reader when the same film is presented at a later time) was slightly better but still imperfect, with kappas of 0.772 for abnormality and 0.662 for parenchymal disease. A study in South Africa, comparing radiographic interpretations of two readers blinded to the patient's clinical characteristics, demonstrated similar interobserver agreement for the findings of abnormality (kappa, 0.47) and parenchymal abnormalities (kappa, 0.75) (21). Another study examined agreement among three chest radiologists who independently evaluated films of patients with suspected miliary TB; they found substantial interobserver agreement (kappa 0.77) for the finding of diffuse nodular opacities (5).

Imperfect reliability of chest radiographic interpretation is not unique to TB; relatively poor agreement among radiologists has been described when assessing the presence of a radiographic infiltrate in subjects with suspected community-acquired pneumonia (22). The imperfect reliability of chest radiograph interpretations is of significant concern given the importance of these interpretations in making clinical decisions. One key finding of our study was that the consensus radiographic interpretation was more highly associated with relapse (OR, 8.97) than the original Study 22 interpretation (OR, 4.97). The reduction in magnitude of the association between a predictor (eg, finding of cavity on CXR) and an outcome (eg, TB relapse) when the predictor is measured with error is known as attenuation (23). Several statistical methods have been proposed to "correct" for attenuation and thereby understand the true magnitude of the relationship between a given predictor variable and an outcome (24). Our validation substudy analysis used one such method (12), and the results suggest that a presumably more accurate reading (ie, the consensus reading) seems to be a much stronger predictor of relapse than the original Study 22 reading. This conclusion has significant implications for both the conduct of TB clinical trials and for public health practice. With regard to clinical trials, centralized reading of all radiographs, preferably by more than one experienced reader blinded to clinical data, may improve the accuracy of detecting key findings such as cavitary disease that correlate with TB outcomes. The latent class analysis performed here provides a measure of independent validation for using a consensus interpretation as the "gold standard." More accurate radiographic interpretation will be more difficult to achieve in the realm of public health practice, but is important to optimize treatment outcomes and resource utilization. Accuracy of radiographic interpretation has been associated with more years of experience and an increasing number of radiographs read annually (25), which suggests that training may improve performance. As Figure 1 demonstrates, even relatively small improvements in performance may lead to significant increases in the ability of radiographic findings to predict clinical outcomes, so even a relatively small effect of training could be worthwhile. Alternatively, given advances in digital technology and teleradiology, centralized reading of TB patients' radiographs by designated experts could have significant potential to improve the accuracy of radiographic evaluation in the public health setting without a major increase in costs. Given the declining incidence of TB in the developed world, with associated decreased TB knowledge and experience among physicians, centralized reading could significantly improve the quality of care provided to TB patients in public health clinics. A caveat is that our results here are derived from statistical modeling of a relatively small subset of radiographs with few relapses. Although the statistical models we used suggest that improving the accuracy of CXR interpretations will increase the ability to predict relapse, validation of these results in a clinical setting is necessary.

Acknowledgments

USPHS/TB Trials Consortium Study 22 was sponsored by the United States Centers for Disease Control and Prevention and was funded in part through a Memorandum of Understanding between the CDC and the Washington, D.C. Veterans Affairs Medical Center. Hoechst Marion Roussel, the manufacturer of rifapentine, provided rifapentine and contributed to the cost of 3 investigator meetings, but did not participate in original or

secondary analysis study design, data collection, data analysis, data interpretation, or writing of the report. Dr. Stout received salary support from the National Institutes of Health (K23AI051409). Dr. Hamilton also received salary support from the National Institutes of Health (K24AI001833).

References

1. Blumberg HM, Burman WJ, Chaisson RE, et al. American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: treatment of tuberculosis. *Am J Respir Crit Care Med*. 2003; 167:603–662. [PubMed: 12588714]
2. 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; 360:528–534. [PubMed: 12241657]
3. Nyboe J. Results of the international study on x-ray classification. *Bull Int Union Tuberc*. 1968; 41:115–124. [PubMed: 5710244]
4. Zellweger JP, Heinzer R, Touray M, et al. Intra-observer and overall agreement in the radiological assessment of tuberculosis. *Int J Tuberc Lung Dis*. 2006; 10:1123–1126. [PubMed: 17044205]
5. Kwong JS, Carignan S, Kang EY, et al. Miliary tuberculosis. Diagnostic accuracy of chest radiography. *Chest*. 1996; 110:339–342. [PubMed: 8697830]
6. Graham S, Das GK, Hidvegi RJ, et al. Chest radiograph abnormalities associated with tuberculosis: reproducibility and yield of active cases. *Int J Tuberc Lung Dis*. 2002; 6:137–142. [PubMed: 11931412]
7. Yerushalmy J, Garland LH, Harkness JT, et al. An evaluation of the role of serial chest roentgenograms in estimating the progress of disease in patients with pulmonary tuberculosis. *Am Rev Tuberc*. 1951; 64:225–248. [PubMed: 14868842]
8. Balabanova Y, Coker R, Fedorin I, et al. Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. *BMJ*. 2005; 331:379–382. [PubMed: 16096305]
9. Dawson-Saunders, B.; Trapp, R. Basic and clinical biostatistics. Englewood Cliffs, NJ: Appleton and Lange; 1994. Estimating and comparing proportions.
10. Benjamin DK, DeLong E, Steinbach WJ. Latent class analysis: an illustrative application for education in the assessment of resident otoscopic skills. *Ambul Pediatr*. 2004; 4:13–17. [PubMed: 14731098]
11. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol*. 1988; 41:923–937. [PubMed: 3054000]
12. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*. 2003; 59:163–171. [PubMed: 12762453]
13. Little, RJA.; Rubin, DB. Statistical analysis with missing data. New York: Wiley; 1987.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
15. Schmidt PG. Open negative cavities. *Tuberkulosearzt*. 1963; 17:137–151. [PubMed: 13992030]
16. Oyama T, Tatsuoka M. Prediction of relapse in pulmonary tuberculosis; an application of discriminant analysis. *Am Rev Tuberc*. 1956; 73:472–484. [PubMed: 13302679]
17. Tripathy SP. Relapse in tuberculosis. *Ind J Tuberc*. 1981; 28:45–57.
18. Aber VR, Nunn AJ. Short term chemotherapy of tuberculosis. Factors affecting relapse following short term chemotherapy. *Bull Int Union Tuberc*. 1978; 53:260–264. [PubMed: 387140]
19. Johnson JL, Okwera A, Vjecha MJ, et al. Risk factors for relapse in human immunodeficiency virus type 1 infected adults with pulmonary tuberculosis. *Int J Tuberc Lung Dis*. 1997; 1:446–453. [PubMed: 9441100]
20. Waaler HT. Description of the study material and organization of the study. *Bull Int Union Tuberc*. 1968; 41:110–114. [PubMed: 5710243]
21. Den Boon S, Bateman ED, Enarson DA, et al. Development and evaluation of a new chest radiograph reading and recording system for epidemiological surveys of tuberculosis and lung disease. *Int J Tuberc Lung Dis*. 2005; 9:1088–1096. [PubMed: 16229219]

22. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators Chest. 1996; 110:343–350.
23. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904; 15:72–101.
24. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. Annu Rev Public Health. 1993; 14:69–93. [PubMed: 8323607]
25. Koppaka, R.; Bock, N. World Health Organization. How reliable is chest radiography?. In: Toman, K.; Frieden, T., editors. Toman's tuberculosis case detection, treatment, and monitoring: questions and answers. Geneva: World Health Organization; 2004. p. 51-60.

Figure 1.

Predicted increase in the magnitude of the relationship between cavity on early chest radiograph and risk of tuberculosis relapse as the sensitivity and specificity of the radiograph interpretation improve. Each curve demonstrates the predicted relationship between the sensitivity of chest radiograph interpretations (as compared to a theoretical “gold standard” consensus reading) and the odds ratio associating cavity and relapse, with specificity held constant. The odds ratio of 12.42 represents the predicted relationship between cavity on early chest radiograph and relapse given perfect agreement with the gold standard (*open square*). The open triangle represents the predicted relationship between cavity on early chest radiograph and relapse given the observed performance characteristics of the original Study 22 radiographic interpretations.

TABLE 1

Agreement among Three Independent, Blinded Readers with Regard to Chest Radiograph Findings in a Subset of Participants in Tuberculosis Trials Consortium Study 22 (99 Subjects, 241 Radiographs)

Finding	% Agreement (kappa value)			Prevalence of finding on consensus interpretation (n = 241)
	Reader1 -Reader2	Reader1 -Reader3	Reader2 -Reader3	
Normal film	93 (0.35)	95 (0.45)	95 (0.45)	94% normal
Bilateral	85 (0.71)	93 (0.86)	87 (0.74)	56% bilateral
Cavitary	84 (0.66)	85 (0.67)	88 (0.73)	34% cavitary
Extension of old lesion(s)	85 (0.70)	87 (0.75)	85 (0.70)	46% with extension
New lesion	82 (0.28)	90 (0.47)	87 (0.41)	12% with new lesions

Readers were blinded to all clinical data (including relapse/failure status) as well as to the chronologic order of the radiographs.

TABLE 2

Sensitivity and Specificity of Blinded Readers for the Finding of Cavity on Chest Radiographs Obtained after 2 Months of Tuberculosis Treatment, as Estimated by Comparison to the Consensus Interpretation and as Estimated by Latent Class Modeling

Reader	<u>Sensitivity for cavity (95% confidence interval)</u>		<u>Specificity for cavity (95% confidence interval)</u>	
	Consensus	Latent class	Consensus	Latent class
Reader 1	0.85 (0.66–0.96)	0.88 (0.66–0.96)	0.90 (0.79–0.97)	0.92 (0.79–0.97)
Reader 2	0.89 (0.71–0.98)	0.92 (0.69–0.98)	0.90 (0.79–0.97)	0.92 (0.79–0.97)
Reader 3	0.96 (0.81–1.0)	0.92 (0.67–0.99)	0.98 (0.90–1.0)	0.96 (0.82–0.99)

TABLE 3

Sensitivity and Specificity of the Original Site Interpretations of Study 22 Chest Radiographs as Compared to the Consensus Interpretation of Three Blinded Readers

Finding	Initial radiograph		2-month radiograph		End of treatment radiograph	
	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Abnormal	1.00 (0.95–1.0)	0.50 (0.01–0.99)	1.00 (0.96–1.0)	0.75 (0.19–0.99)	0.94 (0.88–0.98)	0.75 (0.35–0.97)
Bilateral	0.81 (0.64–0.92)	0.93 (0.76–0.99)	0.89 (0.76–0.96)	0.91 (0.76–0.98)	0.74 (0.60–0.85)	0.98 (0.88–1.0)
Cavitary	0.70 (0.47–0.87)	0.86 (0.71–0.95)	0.81 (0.61–0.93)	0.82 (0.68–0.91)	0.47* (0.28–0.66)	0.93 (0.84–0.98)

* $P = .013$ compared to 2-month sensitivity estimate (Fisher exact test).