Eur Radiol (2013) 23:2459–2468
DOI 10.1007/s00330-013-2840-z

CHEST

The effect of a tuberculosis chest X-ray image reference set on non-expert reader performance

Catriona J. Waitt · Elizabeth C. Joekes · Natasha Jesudason · Peter I. Waitt · Patrick Goodson · Ganizani Likumbo · Samuel Kampondeni · E. Brian Faragher · S. Bertel Squire

Received: 19 November 2012 / Revised: 12 February 2013 / Accepted: 26 February 2013 / Published online: 8 May 2013

Abstract

Objectives In low-resource settings, limitations in diagnostic accuracy of chest X-rays (CXR) for pulmonary tuberculosis (PTB) relate partly to non-expert interpretation. We piloted a TB CXR Image Reference Set (TIRS) to improve non-expert performance in an operational setting in Malawi.

Methods Nineteen doctors and clinical officers read 60 CXR of patients with suspected PTB, at baseline and using TIRS. Two officers also used the CXR Reading and Recording System (CRRS). Correct treatment decisions were assessed against a "gold standard" of mycobacterial culture and expert performance.

Results TIRS significantly increased overall non-expert sensitivity from 67.6 (SD 14.9) to 75.5 (SD 11.1, *P*=0.013), approaching expert values of 84.2 (SD 5.2). Among doctors, correct decisions increased from 60.7 % (SD 7.9) to 67.1 % (SD 8.0, *P*=0.054). Clinical officers increased in sensitivity from 68.0 % (SD 15) to 77.4 % (SD 10.7, *P*=0.056), but decreased in specificity from 55.0 % (SD 23.9) to 40.8 % (SD 10.4, *P*=0.049). Two officers made correct treatment decisions with TIRS in 62.7 %. CRRS training increased this to 67.8 %.

Conclusion Use of a CXR image reference set increased correct decisions by doctors to treat PTB. This tool may provide a low-cost intervention improving non-expert performance, translating into improved clinical care. Further evaluation is warranted.

Electronic supplementary material The online version of this article (doi:10.1007/s00330-013-2840-z) contains supplementary material, which is available to authorized users.

C. J. Waitt
Malawi-Liverpool-Wellcome Clinical Research Programme,
PO Box 30096, Chichiri,
Blantyre 3, Malawi

C. J. Waitt
Department of Molecular and Clinical Pharmacology,
University of Liverpool, Block A, The Waterhouse Buildings,
1-5 Brownlow Street,
Liverpool L69 3GL, UK

E. C. Joekes
Department of Radiology, Royal Liverpool University Hospital,
Prescot Street,
Liverpool L7 8XP, UK

E. C. Joekes · N. Jesudason · E. B. Faragher · S. B. Squire
Liverpool School of Tropical Medicine, Pembroke Place,
Liverpool L3 5QA, UK

P. I. Waitt · P. Goodson · G. Likumbo
Department of Medicine, Queen Elizabeth
Central Hospital, Chichiri,
Blantyre, Malawi

S. Kampondeni
Department of Radiology, Queen Elizabeth
Central Hospital, Chichiri,
Blantyre, Malawi

C. J. Waitt (✉)
The Wolfson Centre for Personalised Medicine,
Department of Pharmacology, University of Liverpool,
2nd Floor, Block A: Waterhouse Buildings, 1-5 Brownlow Street,
Liverpool L69 3GL, UK
e-mail: cwaitt@liv.ac.uk

Key Points

- *Tuberculosis treatment decisions are influenced by CXR findings, despite improved laboratory diagnostics.*
- *In low-resource settings, CXR interpretation is performed largely by non-experts.*
- *We piloted the effect of a simple reference training set of CXRs.*
- *Use of the reference set increased the number of correct treatment decisions. This effect was more marked for doctors than clinical officers.*
- *Further evaluation of this simple training tool is warranted.*

Keywords Radiography · Tuberculosis · Malawi · Sensitivity and Specificity · Teaching

Introduction

Despite the announcement by the World Health Organisation (WHO) in 1993 that tuberculosis (TB) was a 'global emergency' requiring significant investment in both programmatic and research sectors, there were almost 9 million new TB cases globally in 2011, with 1.5 million deaths [1]. Effective treatment relies on a lengthy drug regimen and diagnosis remains challenging. Sputum smear microscopy and culture are established diagnostic standards [1], with chest X-ray (CXR) being called into question due to limited diagnostic accuracy and poor film quality, particularly in low-resource settings. Despite these limitations, many diagnostic algorithms still include CXR [2–5]. Furthermore, in resource-poor settings, where the majority of TB patients are diagnosed, infrastructure is often erratic and unreliable [1]. Smear-microscopy becomes less accurate when throughput is high and depends on the health and presence of key workers [6]. Even as more sensitive molecular tests for pulmonary tuberculosis (PTB), such as GeneXpert, become available, CXR remains necessary in the evaluation of patients with compatible symptoms but negative laboratory results. Therefore, there remains a need to address poor film quality and improve observer performance.

Few investigators have sought to improve CXR interpretation by non-experts in the routine diagnostic process, despite the fact that most of the readers globally are necessarily non-expert. In low-resource settings, clinical care is largely delivered by single-handed, often junior, physicians or clinical officers (COs), who have received training of variable length and quality. In addition, it is well recognised that non-experts show lower diagnostic accuracy than experts [7]. Recruiting COs to undertake screening using CXR showed good results, but required intensive training and supervision [8]. Manuals and short courses are available, some of which advocate standardised radiological reporting, but their effect has not been validated. In addition, retaining complex reporting skills is a challenge, while cost and staffing constraints preclude large-scale access to such training.

As an alternative approach, we hypothesised that an interpretation aid based on comparison of the CXR with a set of reference images could be a more practical, low-cost tool to improve non-expert performance. The objective of this study was to pilot such a tool under operational conditions in Malawi. In addition, we compared the tool with the effect of short course training for standardised reporting.

Methods

Study setting

The study was conducted between April and June 2010, at the Queen Elizabeth Central Hospital (QECH), a tertiary referral hospital in Blantyre, Malawi. Blantyre had an estimated TB incidence of 304/100 000 in 2009 (WHO report 2009—global tuberculosis control). Malawi is among the ten poorest countries worldwide in terms of both GDP and per capita income, and has approximately one doctor per 100,000 of the population. Ethical approval was granted by the College of Medicine Research Ethics Committee, University of Malawi (P.04/05/353) and the Ethics committee of the Liverpool School of Tropical Medicine.

Participants

All COs and junior or middle-grade doctors (DRs) in the Department of Medicine were invited. A CO has completed a 3-year diploma in Clinical Medicine, Surgery and Community Health and a 1-year internship. A junior doctor has 1 year's post-qualification experience and a middle-grade of at least 2. A questionnaire on training and experience was completed (online appendix A). Three radiologists, with at least 5 years' experience in TB CXR participated S.K., E.J., K.I., S.K. being the sole radiologist in Malawi when this study was undertaken. Two COs were selected to attend the Chest Radiograph Reading and Recording System (CRRS) course in May 2010 in South Africa.

CXR test set

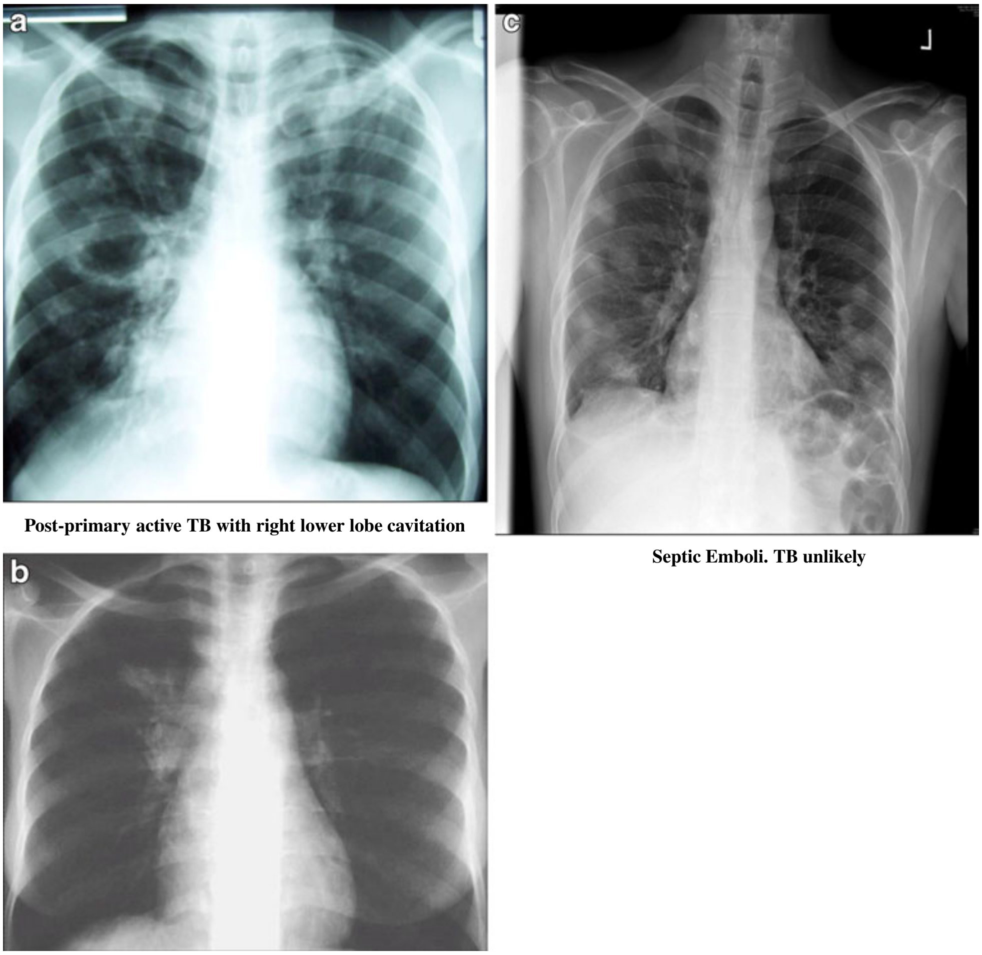
The test set was compiled from new, adult PTB suspects participating in a study investigating early mortality in PTB [9]. Unlike most of the TB cases managed at QECH, CXR and microbiological culture results were available for all. To minimise confounding factors in this pilot study, those with previous or extra-pulmonary disease were excluded, as well as HIV positive patients who had been established on antiretroviral therapy. The principal investigator (C.W.) randomly selected 60 X-rays from smear-positive, culture-positive (23), smear-negative, culture-positive (17) and smear-negative, culture-negative (20) patients. All films were anonymised and digitised [10]. The culture-negative group included six normal X-rays, three films suggestive of PTB and ten X-rays with abnormalities not considered characteristic of PTB. Of these, six had confirmed diagnoses: cardiac failure, metastases, lymphoma, *Pneumocystis jirovecii* pneumonia, Kaposi's sarcoma and pneumococcal pneumonia. The remaining films were suggestive of congestive cardiac failure, primary lung malignancy and chronic obstructive pulmonary disease. HIV status was not taken into account during film selection, but retrospectively documented as 68 %.

Tuberculosis CXR image reference set

Anonymised images were selected from teaching material (E.J.), including both high-quality digital X-rays, as well as X-rays from settings similar to Malawi. A spectrum of PTB appearances in adults was presented in 17 black and white, A4 paper prints of JPEG files. This included two examples of atypical presentation in HIV co-infected patients, e.g. focal lower lobe pneumonia. In addition, examples of a normal CXR, an over-exposed and an under-exposed film were included, as well as seven examples of common diagnoses, not typical of PTB, e.g. cardiac failure and septic emboli. Text was deliberately kept to a minimum, describing only the intended use of the tool, the characteristics of a normal film and diagnosis and likelihood of PTB for each image. Figure 1 illustrates sample images.

Reading of the CXR test set

Readers were aware that films were from PTB suspects, but blinded to other results. They answered two questions: "Would you treat this as PTB? Yes or no?" and: "How certain are you of your decision?" (online appendix B). This format, rather than a detailed radiological film description, was chosen to reflect the clinical setting and force a decision [11]. Reading time was documented. All participants reviewed the set at baseline. Subsequently, the set was re-read by COs and DRs, using the Tuberculosis CXR Image Reference Set (TIRS). Two COs then attended the CRRS course in South Africa and re-read the test set



Post-primary active TB with right lower lobe cavitation

Septic Emboli. TB unlikely

Right hilar adenopathy and right upper lobe bronchopneumonia. Active TB

Fig. 1 Three examples of chest X-rays from the Tuberculosis CXR Image Reference Set (TIRS) booklet. Multiple presentations of active TB were included (**a**, **b**), as well as several common pitfalls in TB diagnosis, for example septic emboli (**c**)

after a 3-week interval on return to Malawi. One week later these two COs re-read the set again using the standardised CRRS proforma (online appendix C). Before every reading the set was reshuffled.

Outcome measures

The primary outcome was the number of correct decisions to treat for PTB, using TIRS, compared with the "gold standard" of mycobacterial culture. Secondary outcomes were the comparison of performance between non-experts and experts, effect of CRRS training, level of certainty of diagnosis and time taken to read.

Data analysis

Results for one culture-negative film were removed after expert readings identified miliary TB (categorised as extrapulmonary TB). Using the remaining 59 films and subsequently the subset of smear-negative, culture-positive films, the following were computed for the treatment decisions of each participant at baseline and for all non-experts using TIRS (taking culture outcome as the "gold standard"): sensitivity, specificity, absolute percentage agreement between treatment decision and culture result, as well as this agreement corrected for chance (kappa statistic). The kappa statistic was calculated for chance agreement with the gold-standard of culture, rather than for inter-observer agreement as is commonly used in imaging studies because our hypothesis related to increasing the number of correct decisions, rather than increasing observer agreement. Participants rated their level of certainty on a four-point scale: 1=0 % sure, 2=up to 30 % sure, 3=30–60 % sure, 4=60–100 % sure. Averages of these statistics are reported for the CO and DR cohorts, both separately and combined, as means and standard deviations; the changes from baseline to TIRS are reported as mean differences with their 95 % confidence intervals (adjusted for clustering within participants). Mean changes in certainty scores were also computed for correct and incorrect diagnoses separately. Finally, the average scores at baseline, with TIRS, and then with CRRS, are reported for the two COs who went to CRRS; no formal statistical comparisons were possible with this small sample. Similarly, results were not considered separately by HIV status, given the low power generated by inclusion of only 19 films from HIV-negative subjects in the test set. Statistical significance was set at the conventional 5 % level for all analyses. Kappa values of>0.207 were considered to be statistically significant.

Results

Participants

The study recruited 24 participants: 11 COs, 10 DRs and 3 radiologists. Two DRs failed to complete all readings and were excluded from analysis. DRs had significantly more self-reported training in CXR interpretation than COs (88 % vs 63 %, P=0.02).

Accuracy of decision to treat

Changes in mean percentage of agreement with the culture gold standard and in agreement corrected for chance for non-experts at baseline and with TIRS, and subset analysis for the smear-negative, culture-positive CXRs, are summarised in Tables 1 and 2. Comparison with expert agreement, sensitivity and specificity are shown in Figs. 2 and 3.

Table 1 Agreement between decision to treat and culture gold standard at baseline and with TIRS—all films

Rater grade		Mean (SD)		Difference (95 % confidence interval)[a] [P value]		
		Baseline	With TIRS			
COs (n=11)	Agreement (%)	63.8 (7.4)	65.7 (5.8)	1.9	(−2.5 to 6.4)	[0.352]
	Kappa	0.210 (0.159)	0.191 (0.100)	−0.019	(−0.102 to 0.064)	[0.622]
DRs (n=8)	Agreement	60.7 (7.9)	67.1 (8.0)	6.4	(−0.2 to 12.9)	[0.054]
	Kappa	0.141 (0.102)	0.276 (0.141)	0.135	(−0.016 to 0.286)	[0.073]
COs+DRs (n=19)	Agreement	62.5 (7.6)	66.3 (6.6)	3.8	(0.3 to 7.3)	[0.035]
	Kappa	0.181 (0.139)	0.227 (0.123)	0.046	(−0.034 to 0.125)	[0.241]
Radiologists[b] (n=3)	Agreement	67.8 (8.9)			–	
	Kappa	0.347 (0.040)			–	

COs clinical officers *DRs* doctors

[a] 95 % confidence interval and P values adjusted for clustering within raters

[b] Only baseline readings done by radiologists

Table 2 Agreement between decision to treat and culture gold standard at baseline and with TIRS—smear-negative, culture-positive subset of 17 films

Rater grade		Mean (SD)		Difference (95 % confidence interval)[a] [*P* value]		
		Baseline	With TIRS			
COs (*n*=11)	Agreement (%)	60.8 (9.0)	56.0 (5.5)	−4.8	(−10.7 to 1.1)	[0.099]
	Kappa	0.228 (0.159)	0.156 (0.083)	−0.072	(−0.184 to 0.040)	[0.183]
DRs (*n*=8)	Agreement	56.6 (6.9)	58.9 (8.3)	2.3	(−8.4 to 13.0)	[0.623]
	Kappa	0.154 (0.135)	0.191 (0.163)	0.037	(−0.160 to 0.234)	[0.672]
COs+DRs (*n*=19)	Agreement	59.0 (8.2)	57.2 (6.8)	−1.8	(−7.0 to 3.5)	[0.482]
	Kappa	0.197 (0.150)	0.171 (0.120)	−0.026	(−0.123 to 0.070)	[0.578]
Radiologists[b] (*n*=3)	Agreement	60.5 (2.7)			–	
	Kappa	0.238 (0.039)			–	

COs clinical officers *DRs* doctors

[a] 95 % confidence interval and *P* values adjusted for clustering within raters

[b] Only baseline readings done by radiologists

Non-expert accuracy using TIRS, compared with baseline

For non-experts combined, TIRS significantly increased the mean percentage of correct decisions to treat from 62.5 (SD 7.6) to 66.3 (SD 6.6, *P*=0.035). Agreement corrected for chance improved, but not significantly (kappa 0.181 [95 % CI 0.139] and 0.227 [95 % CI 0.123, *P*=0.241]). Sensitivity increased significantly from 67.6 (SD 14.9) to 75.5 (SD 11.1, *P*=0.013). Specificity did not change. Results for COs and DRs analysed separately show a non-significant increase in percentage of correct decisions for DRs from 60.7 (SD 7.9) to 67.1 (SD 8.0, *P*=0.054), also when corrected for chance, with kappa increasing from 0.141 (95 % CI 0.102) to 0.276 (95 % CI 0.141, *P*=0.073). Sensitivity and specificity of DRs improved, but did not reach significance. For the COs the percentage of correct decisions to treat remained unchanged. Their sensitivity increased from 68.0 (SD 15) to 77.4 (SD 10.7, *P*=0.056), while specificity decreased significantly from 55.0 (SD 23.9) to 40.8 (SD 10.4, *P*=0.049).

In the subset analysis of smear-negative, culture-positive films, TIRS did not change any of the parameters for the non-expert group as a whole, nor for the subgroup of DRs. For COs the percentage of correct decisions did not change, but a non-significant reduction in specificity from 54.1 (SD 23.2) to 40.3 (SD 11.9, *P*=0.051) was noted.

Non-expert accuracy compared with experts

Experts showed a higher mean percentage of correct decisions at baseline than all non-experts combined. However, this difference was not significant: 67.8 (SD 8.9) vs 62.5 (SD 7.6). Specificity was equal at 49.1 (SD 11.0) and 51.7 (SD 20.3). Sensitivity of experts was higher at 84.2 (SD 5.2) vs 67.6 (SD 14.9). Using TIRS, non-expert percentage agreement and specificity did not change significantly. Sensitivity increased to 75.5 (SD 11.1) and approached expert values at 84.2 (SD 5.2). For the subgroups of DRs and COs, all results followed this same pattern.

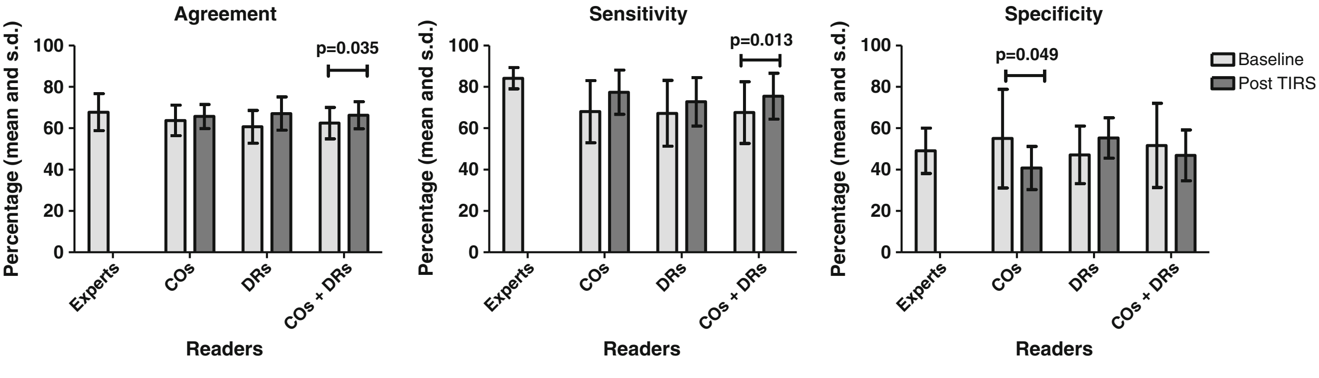


Fig. 2 Agreement with culture gold standard, sensitivity and specificity at baseline and with TIRS—all films

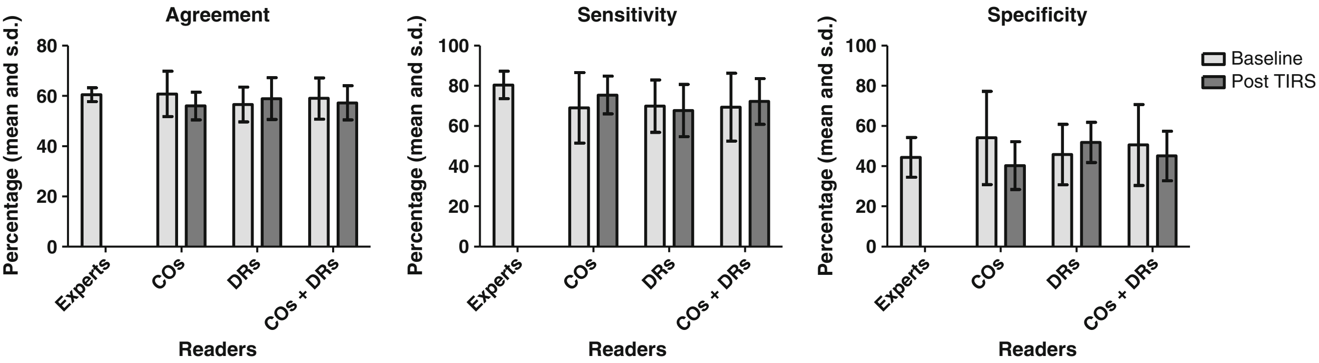


Fig. 3 Agreement with culture gold standard, sensitivity and specificity at baseline and with TIRS—smear-negative, culture-positive subset of 17 films

For the smear-negative subset there were no significant differences between experts and non-experts in mean percentage of agreement with the gold standard or specificity, either at baseline or using TIRS (Fig. 3). Mean sensitivity for experts was higher than for non-experts at 80.4 (SD 6.8) vs 69.4 (SD 16.9), but not significantly different. The same was noted for both DR and CO subgroups.

Clinical officer accuracy after CRRS training

Results for readings after CRRS are presented in Table 3. The small number of CXRs in the smear-negative subset precluded meaningful subgroup analysis for these two readers. Using the standardised CRRS proforma, the average percentage agreement with the gold standard increased from 62.7 (with TIRS) to 71.2, with agreement corrected for chance increasing from κ=0.181 to 0.367 (Table 3). Sensitivity increased from 68.8 to 75.0 and specificity from 50 to 63. Using our question regarding decision to treat, the percentage agreements and kappa increased to 67.8 and 0.256 respectively. Sensitivity increased to 79.8, while specificity decreased to 44.8 %.

Levels of certainty

Certainty scores with TIRS for all non-experts increased significantly from 3.08 (SD 0.78) to 3.38 (SD 0.69, *P*=0.002) for a correct decision and from 2.92 (SD 0.82) to 3.28 (SD 0.76, *P*= 0.002) for an incorrect decision. There was no change in certainty levels for the subgroup of DRs, while COs showed significantly increased confidence for correct and incorrect decisions. A similar increase in scores was noted after CRRS. Expert certainty scores were significantly higher for correct than for incorrect decisions and higher than non-expert scores overall.

Time taken to read

The TIRS increased reading time from 50 s to 70 s per film. The CRRS standardised proforma increased time from 50 s per film to just over 4 min.

Discussion

This pilot shows that TIRS demonstrated a trend towards increasing the number of correct decisions to treat PTB among DRs in the routine clinical setting in Malawi, with improvements in both sensitivity and agreement between CXR interpretation and mycobacterial culture. For COs the outcomes were different, showing no increase in the number of correct decisions, but a trend towards increased sensitivity. The associated loss in their specificity indicates a shift in decision-making from under-diagnosis to over-diagnosis. This difference between the two subgroups might be explained by the fact that DRs had received significantly more formal background training in reading CXRs; the tool may enable recall of prior knowledge, including alternative CXR diagnoses. The latter would also account for their improved specificity, as opposed to the COs' reduction in specificity. For the subset of smear-negative films there was no difference in number of correct decisions, which poses a challenge, as this is a group of patients where CXR is particularly important [2, 3]. This may relate to the high HIV prevalence, where CXR has lower accuracy [12].

Interpretation by COs and DRs overlapped with expert interpretation at baseline, and with TIRS (Fig. 2). Sensitivity was similar to previous reports of non-expert readings from Nepal and Malawi and to expert readings from Kenya and South Africa (Table 4) [3, 12, 19]. Specificity was lower, which is most likely related to high HIV prevalence in our test set and the use of expert reference standards, rather than mycobacterial culture, in previous non-expert reading studies. Standard deviations (SD) for non-expert readers were wide, despite the relatively large number of readers. This reflects the difficulties non-experts encountered. Interestingly, the almost randomly achieved (κ<0.2068) correct number of decisions by non-experts at baseline is very close to expert numbers, and performance for both groups is limited with 61–68 % correct decisions. However, when analysed in more detail, experts appear to take a more informed decision (with higher agreement corrected for chance) and err on the

Table 3 Agreement with culture gold standard, sensitivity, specificity and agreement corrected for chance for two clinical officers (*COA* and *COB*) attending the CXR Reading and Recording System (*CRRS*)

Rater	Baseline				With TIRS				After CRRS (with CRRS proforma)				After CRRS (with question to treat)			
	Agree (%)	Sens. (%)	Spec. (%)	κ	Agree (%)	Sens. (%)	Spec. (%)	κ	Agree (%)	Sens. (%)	Spec. (%)	κ	Agree (%)	Sens. (%)	Spec. (%)	κ
COA	57.6	60.0	52.6	0.115	59.3	65.0	47.4	0.117	69.5	72.5	63.2	0.338	66.1	75.0	47.4	0.224
COB	59.3	62.5	52.6	0.140	66.1	72.5	52.6	0.245	72.9	77.5	63.2	0.396	69.5	84.6	42.1	0.287
Mean	58.5	61.3	52.6	0.128	62.7	68.8	50.0	0.181	71.2	75.0	63.2	0.367	67.8	79.8	44.8	0.256

side of false-positive treatment decisions (higher sensitivity). With TIRS, non-experts and, in particular, DRs not only improve their overall performance, but alter their decision-making to coincide more with expert decision patterns. In the context of inadequate TB case-detection in many developing countries [1], a tendency to over-treat is better for TB control than a tendency to under-treat.

The CRRS training and standardised reporting were developed for prevalence surveys. Good inter-observer agreement between two experts was reported [16]. However, subsequent use in HIV screening showed sensitivity and specificity similar to several other studies using non-standardised reporting (Table 4) [13]. It is advocated for use in the clinical setting, but to our knowledge has not been validated. We observed an increase in the number of correct decisions from 63 % with TIRS to 71 % using the CRRS proforma, with an associated increase in agreement corrected for chance. However, this increase was less marked (63 % to 68 %) when using the question: “Would you treat?” and specificity in particular dropped to baseline levels. As completing the CRRS proforma increased reading time from 1 min to 4 min per film, this poses a challenge in a busy clinical setting

Non-experts are generally aware of their limited diagnostic accuracy and, therefore, we were interested to assess whether TIRS might increase confidence and reduce the desire for second opinions. Although DRs showed improvement in accuracy, confidence did not alter. Confidence of COs did increase, regardless of their decision, in effect providing a false sense of security with a potentially negative effect on referral patterns.

A major strength of this pilot is our rigorous assessment of the impact of TIRS in keeping with current guidelines on reporting of diagnostic accuracy studies [20]. Studies of CXR performance in PTB vary widely in outcome (Table 4). This is partly related to variations in prevalence and imaging characteristics in the population tested, but more often to methodology applied, such as number of readers, choice of reference standard and presentation of results as observer agreement or sensitivity and specificity. This creates confusion in the interpretation of study outcomes and the perceived validity of CXRs. For example, good observer agreement between two experts does not necessarily equate to high diagnostic accuracy against a culture gold standard. In addition, an expert panel is often the only reference standard available, but in PTB is particularly prone to error. As we have shown, our expert readers were incorrect in 30 % of cases. To our knowledge, this pilot is the first assessment of non-expert performance against a gold standard of culture. In addition, we corrected for chance agreement with the gold standard and for clustering within raters. This resulted in a rigorous assessment of effect. For example, COs apparently improved with TIRS, but once

Table 4 Previous studies on accuracy of chest X-ray (CXR) interpretation in tuberculosis

Author	Year	Setting	Film sets[a]	Sensitivity % (95 %CI)	Specificity % (95 %CI)	Inter-observer kappa (95 % CI)	Readers[b]	CRRS	Reference standard
Den Boon [16]	2005	South Africa	Prevalence study			0.69 (0.64–0.74) for PTB 0.47 (0.42–0.53) for normal	Expert (1)	√	Single expert
Agizew [13]	2010	Botswana	Screening in HIV			“no disagreement”	Expert (2)	√	Culture/follow-up[c]
Dawson [15]	2010	South Africa	Screening in HIV	68 (54–79)	53 (45–61)	0.61 (0.40–0.83)	Expert (2)	√	Culture
Cain [3]	2010	SE Asia	Screening in HIV	65	85		Expert (1)		Culture
Davis [12]	2010	Uganda	HIV positive suspects	78 (66–87)	22 (16–30)		Expert (2)	√	Culture/follow-up
van Cleeff [19]	2005	Kenya	HIV negative suspects	80 (74–85)	67 (62–71)	0.75 (SE 0.037)	Expert (2)		Culture
Nyirenda [18]	1999	Malawi	HIV negative suspects	65–71	71–79		Non-expert (194)		Expert reference panel
Balabanova [14]	2005	Russia	Population screening			0.387 (0.382–0.391)	Expert (101)		Expert reference panel
Kumar [17]	2005	Nepal	HIV negative suspects	60	72	0.17 (0–0.38)	Non-expert (21)		Expert reference panel
Zellweger [7]	2006	Switzerland	Immigrant screening			0.846 (SE 0.029) 0.557 (SE 0.109)	Expert (2) Expert (2) and non-expert (1)		None

[a] Population from which the film sets were obtained

[b] Experts include pulmonologists, pulmonary tuberculosis (PTB) specialists and radiologists. Non-experts include all other levels of CXR readers

[c] PTB diagnosis was presumptive in 55 %

corrected for chance, may actually still be guessing. Similarly, using a large number of non-expert readers and dividing them into subgroups of DRs and COs with information on levels of training helped to explain the different effect of TIRS on each subgroup.

Several limitations are present. The results for the CRRS readings should be viewed with caution as only two COs could attend. This highlights the prohibitive expenses in settings similar to Malawi of attending such courses. Using a simulation film set does not fully reflect clinical practice. Similarly, excluding patients with a previous history of TB may have influenced results. Paper prints of X-rays limit image resolution and visibility of small nodules, potentially important for PTB diagnosis. However, most films in low resource settings are of limited quality, with similar limitations in resolution. Further validation, in a clinical setting and including cases with previous TB, will be required. In addition, we used the same film set for all readings, which may have biased results, despite films being re-shuffled. In the culture-negative subset not all diagnoses could be confirmed, owing to limited resources, and culture-negative PTB may have been present.

In the population tested, it is arguable whether improving non-expert performance is required, as experts performed only marginally better. On the other hand, although improvements were modest, some promise for the DRs was noted, even in this population. Evaluation of a larger number of non-expert clinicians in a range of populations (e.g. low HIV prevalence/screening in people living with HIV) may reveal superior results. If so, low cost and simplicity are strengths that may justify implementation, even if benefits are small. The lack of effect in the smear-negative group is a limitation, but when access to laboratory tests is limited, the tool may be helpful. Until fast turn-around molecular or microbiological diagnosis of PTB is universally available, CXR will still be used across the globe by non-expert readers to inform treatment decisions. We suggest that low-cost, easily delivered interventions aimed at improving CXR interpretation will have greater overall impact than more expensive and time consuming options such as training courses or teleradiology.

In conclusion, a pulmonary tuberculosis CXR image reference set increased the number of correct decisions to treat pulmonary tuberculosis by non-experts in the operational setting in Malawi. This effect was more marked for doctors than clinical officers. Further evaluation of this tool in clinical practice may provide a validated, simple, low-cost intervention to improve non-expert reader performance.

Acknowledgments We are grateful to all clinical officers and doctors who took time out of their busy daily practice to participate and to Dr Klaus Irion, radiologist, for performing the third expert reading. This study was funded as a Liverpool School of Tropical Medicine MSc dissertation project (NJ). Additional funding to support clinical officer attendance at CRRS in South Africa was kindly provided by Professor Derek Gould, Department of Radiology, Royal Liverpool and Broadgreen University Hospital NHS Trust, from an NIHR senior investigator award. CJW was supported by a Wellcome Trust Training Fellowship in Clinical Tropical Medicine.

References

1. World Health Organization (2011) Global tuberculosis control 2011. World Health Organization, Geneva. Available via http://www.who.int/tb/publications/global_report/en/. Accessed July 2012
2. World Health Organization (2007) Improving the diagnosis and treatment of smear-negative pulmonary and extrapulmonary tuberculosis among adults and adolescents: recommendations for HIV-prevalent and resource-constrained settings. World Health Organization, Geneva. Available via http://whqlibdoc.who.int/hq/2007/WHO_HTM_TB_2007.379_eng.pdf. Accessed July 2012
3. Cain KP, McCarthy KD, Heilig CM et al (2010) An algorithm for tuberculosis screening and diagnosis in people with HIV. N Engl J Med 362:707–716
4. Tamhane A, Chheng P, Dobbs T, Mak S, Sar B, Kimerling ME (2009) Predictors of smear-negative pulmonary tuberculosis in HIV-infected patients, Battambang, Cambodia. Int J Tuberc Lung Dis 13:347–354
5. Hanifa Y, Fielding KL, Charalambous S et al (2012) Tuberculosis among adults starting antiretroviral therapy in South Africa: the need for routine case finding. Int J Tuberc Lung Dis 16:1252–1259
6. Mundy CJ, Harries AD, Banerjee A, Salaniponi FM, Gilks CF, Squire SB (2002) Quality assessment of sputum transportation, smear preparation and AFB microscopy in a rural district in Malawi. Int J Tuberc Lung Dis 6:47–54
7. Zellweger JP, Heinzer R, Touray M, Vidondo B, Altpeter E (2006) Intra-observer and overall agreement in the radiological assessment of tuberculosis. Int J Tuberc Lung Dis 10:1123–1126
8. Hoog AH, Meme HK, van Deutekom H et al (2011) High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. Int J Tuberc Lung Dis 15:1308–1314
9. Waitt CJ, Peter KBN, White SA et al (2011) Early deaths during tuberculosis treatment are associated with depressed innate responses, bacterial infection, and tuberculosis progression. J Infect Dis 204:358–362
10. Szot A, Jacobson FL, Munn S et al (2004) Diagnostic accuracy of chest X-rays acquired using a digital camera for low-cost teleradiology. Int J Med Inform Feb 73:65–73
11. Potchen EJ (2006) Measuring observer performance in chest radiology: some experiences. J Am Coll Radiol 3:423–432
12. Davis JL, Worodria W, Kisembo H et al (2010) Clinical and radiographic factors do not accurately diagnose smear-negative tuberculosis in HIV-infected inpatients in Uganda: a cross-sectional study. PLoS One 5:e9859
13. Agizew T, Bachhuber MA, Nyirenda S et al (2010) Association of chest radiographic abnormalities with tuberculosis disease in asymptomatic HIV-infected adults. Int J Tuberc Lung Dis 14:324–331
14. Balabanova Y, Coker R, Fedorin I et al (2005) Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. BMJ 331:379–382

15. Dawson R, Masuka P, Edwards DJ et al (2010) Chest radiograph reading and recording system: evaluation for tuberculosis screening in patients with advanced HIV. Int J Tuberc Lung Dis 14:52–58
16. Den Boon S, Bateman ED, Enarson et al (2005) 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 9:1088–1096
17. Kumar N, Bhargava SK, Agrawal CS, George K, Karki P, Baral D (2005) Chest radiographs and their reliability in the diagnosis of tuberculosis. JNMA J Nepal Med Assoc 44:138–142
18. Nyirenda TE, Harries AD, Banerjee A, Salaniponi FM (1999) Accuracy of chest radiograph diagnosis for smear-negative pulmonary tuberculosis suspects by hospital clinical staff in Malawi. Trop Doct 29:219–220
19. van Cleeff MR, Kivihya-Ndugga LE, Meme H, Odhiambo JA, Klatser PR (2005) The role and performance of chest X-ray for the diagnosis of tuberculosis: a cost-effectiveness analysis in Nairobi, Kenya. BMC Infect Dis 5:111
20. Bossuyt PM, Reitsma JB, Bruns DE et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. AJR Am J Roentgenol 181:51–55