Aus dem Centrum für Chronische Immundefizienz (CCI)
der Albert-Ludwigs-Universität Freiburg i. Br.

Diagnostic accuracy of Computer-Aided Detection and a Scoring System for Pulmonary Tuberculosis in Chest Radiographs: A Validation Study from Sub-Saharan Africa

INAUGURAL-DISSERTATION

zur
Erlangung des Medizinischen Doktorgrades
der Medizinischen Fakultät der
Albert-Ludwigs-Universität
Freiburg i.Br.

Vorgelegt 2016
von Marianne Breuninger
geboren in Offenburg

Swiss TPH
Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut

Dekanin	Prof. Dr. Kerstin Krieglstein
1. Gutachter	Prof. Dr. Dirk Wagner
2. Gutachter	PD Dr. Markus Hufnagel
Jahr der Promotion	2016

Ein wesentlicher Teil dieser Arbeit wurde in der folgenden Publikation veröffentlicht:

Breuninger M, van Ginneken B, Philipsen RH, Mhimbira F, Hella JJ, Lwilla F, van den Hombergh J, Ross A, Jugheli L, Wagner D, Reither K. (2014).

Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa.

PLoS One

Ein weiterer Teil wurde als Poster präsentiert:

Breuninger, M, van den Hombergh J, Dharsee J, Jugheli L, Hella JJ, Wagner D, Reither K.

Tanzanian X-ray score for the detection of active pulmonary TB on chest radiographs: a comparison with subjective assessment.

Oral Poster Presentation at the 45th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union) (2014).

Zudem entstand im Rahmen der beschriebenen Studie folgende Publikation:

Melendez, J, van Ginneken B, Maduskar P, Philipsen, RH, Reither, K, **Breuninger, M**, Adetifa, IMO, Maane, R, Ayles, H and Sanchez, CI (2015).

A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays.

IEEE Transactions on Medical Imaging 34:179–192.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AFB	acid-fast bacilli
A_z	area under the receiver operating characteristic curve
CAD	computer-aided diagnosis
CE	Conformité Européenne
CI	confidence interval
CRRS	Chest Radiograph Reading and Recording System
CXR	chest X-ray
DIAG	Diagnostic Image Analysis Group at Radboud University Medical Center, Nijmegen, The Netherlands
DST	drug-susceptibility testing
HIV	human immunodeficiency virus
IHI	Ifakara Health Institute
M.D.	medical doctor
M.tb	*Mycobacterium tuberculosis*
NLR	negative likelihood ratio
NPV	negative predictive value
NTM	non-tuberculous mycobacteria
NTP	National Tuberculosis Program
PLR	positive likelihood ratio
p_o	overall percentage agreement

PPV	positive predictive value
p_s	proportions of specific agreement
PTB	pulmonary tuberculosis
ROC	receiver operating characteristic
sens.	sensitivity
spec.	specificity
SSM	sputum smear microscopy
TB	tuberculosis
TIRS	TB CXR Image Reference Set
TXS	Tanzanian Chest X-ray Score
WHO	World Health Organization
ZN	Ziehl-Neelsen
κ_w	weighted κ agreement

Zusammenfassung

Hintergrund: Die fehlende Übereinstimmung zwischen Untersuchern, eine Vielfalt an Befundungsmethoden, sowie der Mangel an erfahrenen Untersuchern in ressourcenschwachen Ländern limitieren das Potential des Thoraxröntgens als Diagnose- und Screening-Test für die pulmonale Tuberkulose (PTB). Zwei Ansätze diese Probleme zu lösen wurden evaluiert: eine computergestützte Auswertung (CAD4TB) sowie ein strukturierter Befundungsbogen mit Score-Funktion (TXS).

Methoden: CAD4TB und TXS wurden an Thorax-Röntgenbildern von Patienten mit PTB-verdächtigen Symptomen in Tansania getestet. Der kulturelle Nachweis von *Mycobacteriuum tuberculosis* (M.tb) galt als Referenzstandard. Die Röntgenbilder wurden von zwei Experten und einem *Clinical Officer* ausgewertet. Die Sensitivität und Spezifität von CAD4TB und TXS wurden mittels ROC-Kurven dargestellt und mit dem konventionellen Befund der Untersucher verglichen. Die Übereinstimmung zwischen TXS-Befund und konventionellem Befund, zwischen Befundungen verschiedener Untersucher, sowie zwischen wiederholten Befundungen durch denselben Untersucher wurde mittels gewichteten Kappa-Koeffizienten (κ_w) berechnet.

Ergebnisse: 193 (22%) der 861 Studienteilnehmer waren kultur-positiv für M.tb. CAD4TB erzielte eine Fläche unter der ROC-Kurve von 0.84 (95% CI 0.80-0.88) für die Erkennung von PTB-Patienten. CAD4TB diagnostizierte PTB zutreffender bei Mikroskopie-positiven und HIV-negativen Patienten gegenüber Mikroskopie-negativen bzw. HIV-positiven Patienten ($p<0.01$). CAD4TB übertraf den *Clinical Officer*, aber erreichte nicht die Treffsicherheit der Experten für die Erkennung tuberkulosespezifischer Auffälligkeiten ($p\leq0.03$). Die Experten erzielten eine substantielle Übereinstimmung (κ_w=0.67) mit Benutzung des TXS als strukturierten Befundungsbogen. Die hohe Übereinstimmung zwischen dem konventionellen und TXS-Befund der Experten (κ_w=0.80/0.79) bezeugen die Validität der Score-Funktion. Die Score-Funktion des TXS verbesserte jedoch weder die diagnostische Treffsicherheit noch die Übereinstimmung der Untersucher und war insbesondere von Nachteil für den *Clinical Officer*.

Schlussfolgerungen: CAD4TB konnte die Röntgenbilder von Patienten mit pulmonaler Tuberkulose treffsicher und reproduzierbar von denen symptomatischer Kontrollpatienten unterscheiden. Experten ist der deskriptive Teil des TXS ein nützlicher Befundungsstandard. Durch die Score-Funktion des TXS lässt sich ihre Auswertung des Röntgenbildes nachvollziehen. Für wenig erfahrene Untersucher ist der TXS nicht geeignet.

Abstract

Background: Chest radiography to diagnose and screen for pulmonary tuberculosis (PTB) has limitations, especially due to inter-reader variability, absence of reporting standards and a lack of experienced readers in resource-constrained settings. We evaluated two efforts to overcome the known drawbacks: a computer-aided diagnosis system (CAD4TB) and a structured reporting and scoring form (TXS).

Methods: CAD4TB and TXS performance was assessed on chest radiographs (CXRs) of patients with symptoms suggestive of PTB in Tanzania. Culture-positive PTB was used as the reference standard. Chest radiographs were read by the software and three human readers, two expert readers and one clinical officer. The sensitivity and specificity of CAD4TB and the TXS was depicted using receiver operating characteristic (ROC) curves and results were compared with manual conclusions of human readers. Agreement between TXS and manual conclusion, inter- and intra-reader agreement was calculated as weighted kappa agreement (κ_w).

Results: Of 861 study participants, 193 (22%) were culture-positive for *Mycobacterium tuberculosis*. The area under the ROC curve of CAD4TB for the detection of culture-positive PTB was 0.84 (95% CI 0.80-0.88). CAD4TB detected PTB more accurate in smear-positive over smear-negative and in HIV-negative compared to HIV-positive individuals ($p<0.01$). CAD4TB outperformed the clinical officer, but did not reach the accuracy of the expert readers for tuberculosis specific reading thresholds ($p\leq0.03$). Inter-expert agreement was substantial (κ_w=0.67) for the use of the TXS as mere structured reporting form. High values of agreement (κ_w=0.80 / 0.79) between the experts' manual and TXS conclusion support the validity of the score. However, the scoring function of the TXS did enhance neither the diagnostic accuracy nor the reproducibility of the human readers. The clinical officer's accuracy for TB consistent findings and intra-reader agreement deteriorated (κ_w=0.35 to 0.20) with the use of the score.

Conclusion: CAD4TB accurately distinguished between the CXRs of culture-positive TB cases and symptomatic controls without the need for trained reading personnel. The descriptive part of the TXS presents a valuable reporting standard for expert readers and its scoring function makes their CXR interpretation transparent. However, in its current form, its use as a score is of no benefit for expert readers and not recommended for non-expert readers.

1 Introduction

1.1 The burden of tuberculosis

Tuberculosis (TB), an infectious disease caused by *Mycobacterium tuberculosis* (M.tb), remains a major global health problem. It is curable and preventable. Nevertheless, an estimated 9.6 million people fell sick with and 1.5 million died from the disease in 2014 (WHO 2015a). Tuberculosis ranked 11th among the top causes of global years of life lost in 2013 (GBD 2013 Mortality and Causes of Death Collaborators 2014). African countries are particularly affected. They accounted for 28% of the world's cases in 2014 and relative to their population for the highest incidence (281 cases / 100.000 population) and prevalence (330 cases / 100.000 population) worldwide (WHO 2015a). The financial burden of tuberculosis for their economies and health care systems is enormous (Kim et al. 2002, WHO 2015a) and costs for the individual often catastrophic (Ukwaja et al. 2012, Tanimura et al. 2014). Tuberculosis affects people in their economically most productive age (Murray et al. 2014). Half of the total costs impend before the diagnosis (Tanimura et al. 2014).

Drug-sensitive tuberculosis is treated with a standard antibiotic regimen that is highly effective and generally well tolerated. Administered in time, it can reduce mortality from 45% to 2.5% in HIV-negative and from 80% to 8-14% in HIV-positive co-infected individuals (Tiemersma et al. 2011, Field et al. 2014, Odone et al. 2014, WHO 2015b). TB treatment is considered as one of the most cost-effective health care interventions (Dye & Floyd 2006). Incomplete and poor quality treatment has selected resistant strains resulting in 5% multidrug-resistant cases of TB in 2014 (WHO 2015a). Their recommended antibiotic regimen is toxic, poorly tolerated, longer in duration (up to 24 months), less potent, much more expensive and misses the evidence of randomized controlled trials (WHO 2013a, Zumla et al. 2013, Dheda et al. 2015). Drug-resistant TB threatens the control of the disease worldwide.

Tuberculosis is a communicable disease. *M. tuberculosis*, an airborne pathogen, is transmitted in droplet nuclei coughed up by patients with active TB. A small number of germs inhaled is enough to become infected. There is no reliable vaccine. Whereas effective treatment renders patients non-contagious almost immediately (Riley et al. 1962), patients that are not treated can infect up to 10 persons a year (WHO 2014a). An estimated 2-3 billion people, that is one third of the world's population, is infected with *M. tuberculosis* (WHO 2015a). Their lifetime risk to develop active disease is 5-15% (WHO 2015a), with a higher probability among people with a compromised immune system, people who are malnourished, suffer from poverty, diabetes, end-stage renal disease, silicosis or who use tobacco (Dheda et al. 2015). HIV co-infected individuals are at 20-40-fold risk to progress to active disease and even on antiretroviral therapy the incidence of tuberculosis stays five-fold higher (Dheda et al. 2015). Tuberculosis in turn accelerates the course of HIV disease (Reid & Shah 2009). In the majority of cases (~85%) the disease manifests as necrotizing granulomatous inflammation of the lungs (pulmonary tuberculosis), but almost any other part of the body can be affected (Dheda et al. 2015). Extrapulmonary tuberculosis (EPTB) is much more common among HIV-positive individuals. The onset of the disease is gradual and possible symptoms like cough, fever, night sweats, weight loss and haemoptysis are rather unspecific.

1.2 TB diagnosis

Initiation of appropriate treatment warrants timely and accurate diagnosis. Tuberculosis is complex to diagnose and few numbers are enough to demonstrate the insufficiency of existing diagnostic methods in keeping pace with the global TB epidemic (figure 1). In 2014, 5.2 million cases of incident pulmonary TB (PTB) were notified to National Tuberculosis Programmes (NTP) worldwide, but only 3 million (58%) received a bacteriological confirmation of their diagnosis by either sputum smear microscopy, culture or the molecular Xpert MTB/RIF (Cepheid Inc.) test (WHO 2015a). For 2.2 million cases of PTB, diagnosis was based on clinical, radiological or histological suspicion (WHO 2015a). A usually long delay from the onset of symptoms to diagnosis and treatment due to both patient and health system reasons (Storla et al. 2008, Sreeramareddy et al. 2009) favours the spread of the disease and economic hardship. Another estimated 3.6 million TB patients were missed at all by NTPs (WHO 2015a). Hence, in more than one third of all TB patients, under-diagnosis and under-reporting result in an unknown quality of care, uninterrupted transmission of the disease and preventable deaths (WHO 2014a).

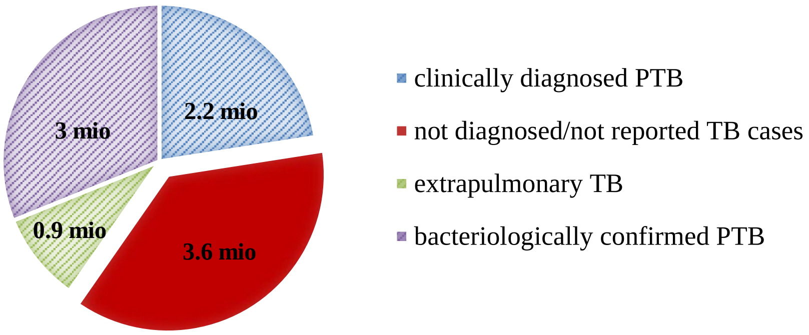


Figure 1 Tuberculosis case detection gap (WHO 2015a)

TB diagnostics are not mutually exclusive. Their suitability and placement in screening and diagnostic algorithms depend much on prevalence, resources and level of care. Three methods for a bacteriological confirmation and definite diagnosis of the disease exist. Their characteristics are summarised in table 1.

Sputum culture is widely regarded as the reference standard to detect active pulmonary TB. It has high sensitivity and specificity (Casal et al. 1997, Asmar & Drancourt 2015), allows for specification of mycobacteria and drug susceptibility testing (DST). The long duration between sample processing and result and the high level of biosafety infrastructure required to safely handle samples are major drawbacks for this method and barriers for its scale-up beyond central laboratory level in resource-limited settings with a high case load (Boyle & Pai 2014).

Most high-burden countries still rely on direct sputum smear microscopy (SSM) as primary method for the diagnosis of active TB (Denkinger et al. 2013). Sputum samples are stained with either colorimetric or fluorescent dye to visualize the presence of mycobacteria under the microscope (Boyle & Pai 2014). This method is highly specific for the detection of acid-fast bacilli, but cannot differentiate between M.tb and non-tuberculous mycobacteria (NTM), between viable and dead organisms or drug-susceptible and drug-resistant strains (WHO 2015c). SSM is low-cost and simple, allows for rapid results and monitoring of treatment progress, requires minimal equipment and biosafety infrastructure (Boyle & Pai 2014). Its major disadvantage is the low sensitivity at an average of 50%, ranging between 20-80% (Boyle & Pai 2014) depending on the bacillary load in the sputum sample and reader skills. This seriously limits its usefulness in children and HIV-positive co-infected individuals, who typically present with paucibacillary disease.

The nucleic acid amplification test Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA) combines the advantages of a rapid diagnosis and high diagnostic accuracy. It detects M.tb and resistance to rifampicin with high sensitivity and specificity in less than 2 hours. It is a fully automated, closed real-time polymerase chain reaction system for the multi-disease Gene Xpert platform (Cepheid). Its operation requires minimal training and biosafety equipment, which makes this technology suitable for regional and district levels of care. Its implementation in peripheral settings is impeded by the operational premises of a stable power supply (not to interrupt testing procedure and lose results), an ambient temperature of < 30°C, security against theft, reliable supply chain management and adequate storage of cartridges (Niemz & Boyle 2012, Denkinger et al. 2013, Weyer et al. 2013).

	average initial capital [1]	**average running costs/ sample** [1]	**through-put**	**sensitivity [%]**	**specificity [%]**	**DST** [2]	**treat-ment moni-toring**	**availability in Tanzania** population = 52 mio. people	**time to result**
(conventional) **culture** (+ DST [2])	1,400,000 (new laboratory) 300,000 (established laboratory)	18.5 [3]	depends on system capacity, up to 8000/year	BacTec MGIT 960 89.4 Lowenstein /Jensen 74	100	yes	yes	4 (1 DST)	liquid media: 10-14 days solid media: 4-6 weeks
smear microscopy - conventional light microscopy (Ziehl-Neelsen stain) - (LED-)fluorescent microscopy	1,500 / microscope	1.77	max. - 25-30/day /fully trained microscopist - 50-60/day	20-80 +10	~98	no	yes	945 (1.8/100,000)	30 min
GeneXpert unit	17,000 / 4 module unit [4]	9.98 [3]	max. 16-20/day	88 (all) 98 (sm+) 68 (sm-)	99	yes [5]	no	59	90 min
digital X-ray unit	100,000 – 180,000	1.5	300/day	87 highly dependent on reader	89 highly dependent on reader	no	no/yes [6]	2	1 min

Table 1 Overview of costs and accuracy of TB diagnostics

[1] Prices in US$, [2] DST = drug-susceptibility testing, [3] price for culture without DST, [4] concessional price for public sectors in eligible countries, [5] drug-susceptibility testing for rifampicin, [6] CXR can be used in smear and culture negative patients to monitor treatment. Data from (Casal et al. 1997, WHO 2015c, Lu et al. 2013, Pantoja et al. 2013, WHO 2013b, Boyle & Pai 2014, Kik et al. 2014, FIND 2015, International Health Partners US 2015, Philipsen, Sánchez, et al. 2015, Asmar & Drancourt 2015)

1.3 Chest radiography & TB

1.3.1 The role of chest radiography in TB diagnosis

Chest radiography constitutes the common first step in the evaluation of patients with pulmonary symptoms in industrialised parts of the world and forms an integral part of diagnostic algorithms in NTP guidelines of most high-burden countries (Pande et al. 2015). A positive chest X-ray (CXR) pre-selects individuals at highest risk for TB for confirmatory testing by Xpert MTB/RIF or culture and supports clinical TB diagnosis where microbiological diagnosis beyond sputum smear microscopy is not feasible. The WHO guidelines for systematic screening for active TB among certain risk groups recommend the use of chest radiography, if available, as first or second screening step (WHO 2013c).

The effective radiation dose of a chest radiograph in adults is comparable to ten days of natural background radiation (The Radiological Society of North America 2015). Conventional (film-based) radiography is still the most common technology in nearly all high burden TB countries (Pande et al. 2015). Its numerous drawbacks limit the potential role of chest radiography in TB screening and diagnosis. The operation is complex with the need for trained personnel, a dark room, supply of films and processing chemicals. Cumulative costs are high and consistent quality assurance of images difficult to maintain (Zennaro et al. 2013). Digital radiography by contrast entails instantly available, high-quality images, lower exposure to radiation, facilitation of electronic storage and transmission of images together with the possibility to use image-processing techniques and computer-aided diagnosis software. Initial investment costs are higher, but prove efficient as running costs are low (Muto et al. 2011) and less trained staff is needed. Innovative digital radiology solutions tailored for low- and middle-income countries aim to make radiological diagnostic accessible, scalable and self-sustainable worldwide (FIND 2015, GlobalDiagnostiX 2015). In March 2015, the GlobalDiagnostiX project presented the prototype of a digital radiography device, which is compact, robust and operable with an alternative power back up, yet at a tenth of the cost of existing equipment and without compromise in image quality (GlobalDiagnostiX 2015). These efforts are urgently needed as almost half of all X-ray machines in developing countries are estimated to be broken (Perry & Malkin 2011) contributing to a lack of access to diagnostic imaging for up to two third of the world’s population (Maru et al. 2010, Pan

American Health Organization 2012). The number of radiologists in the public service of many high burden countries is vanishingly low (Coulborn et al. 2012) and most images are read by non-experts.

1.3.2 The diagnostic accuracy of chest radiography

As a rapid examination technique, which can be interpreted on-site, chest radiography has the potential to shorten diagnostic delays considerably. Together with its high sensitivity**,** it qualifies as an efficient triage test. Readers who focus on TB related abnormalities can identify 87% (95% CI 79-95%) of PTB patients and reach up to 98% (95% CI 95-100%) if they consider any abnormality (Hoog et al. 2014). The modest specificity of 89% (95% CI 87-92%) for TB consistent abnormalities and 75% (95% CI 72-79%) for any abnormality precludes the use of chest radiography as a stand-alone diagnosis (Hoog et al. 2014). Upper lobe infiltrates, fibrosis and cavitary lesions are typical findings in HIV-negative TB patients, but no conclusive proof of the disease. Disease representation in HIV co-infected individuals depends much on their immune status and ranges from mid/lower zone infiltrates, miliary disease, lymphadenopathy and pleural effusions to a normal radiographic appearance in 7-46% of patients (Chamie et al. 2010, Padmapriyadarsini et al. 2013, Swindells et al. 2013). As a result, accuracy of CXR in this particularly vulnerable population is reduced to sensitivity and specificity values between 48-72% and 53-81%, respectively (Dawson et al. 2010, Padmapriyadarsini et al. 2013, Swindells et al. 2013). Computed tomography has a higher sensitivity for the detection of early parenchymal lesions and mediastinal lymph node enlargement as well as for the evaluation of disease activity (Lange & Mori 2010). Non-availability limits its potential benefits in the diagnosis of TB: in Tanzania, a typical sub-Saharan African country, there are no more than six computed tomography scanners for a population of 52,000,000 people (WHO 2014b).

1.3.3 Absence of reporting standards and reader variability: problem and possible solutions

A lack of consistency in how results are reported and high levels of inter- and intra-reader variability (Koppaka & Bock 2004) have been persistent matters of concern. Approaches to promote standardised reporting and to enhance reproducibility, especially among non-expert readers, include handbooks (Tuberculosis Coalition for Technical

Assistance (TBCTA) 2010), training courses (Tanzanian Ministry of Health and Social Welfare et al. 2010), the development of structured reading and reporting systems (Den Boon et al. 2005) and the compilation of a reference image set (Waitt et al. 2013). While some of these attempts have proved useful (Den Boon et al. 2005, Waitt et al. 2013), others are yet to be validated.

The complexity of the interpretation code (Graham et al. 2002, Zellweger et al. 2006) and image quality affect the result. Different readers are also influenced by experience and professional training (Balabanova et al. 2005, Zellweger et al. 2006, Steiner et al. 2015) and momentary factors like distraction, focus and tiredness. The use of simplified codes in the radiological screening of immigrants from high- to low-incidence countries has yielded moderate (Graham et al. 2002) and substantial agreement among readers of different experience levels and almost perfect agreement among expert readers (Zellweger et al. 2006).

The International Labour Organization addressed similar problems in the interpretation of chest radiographs for occupational lung diseases in publishing the International Classification of Radiographs of Pneumoconiosis together with a reference set of images and accreditation system for readers (International Labour Organization 2011). Inspired by the benefits of a simple, reproducible and systematic reporting system, researchers from the University of Cape Town developed the Chest Radiograph Reading and Recording System (CRRS) (Den Boon et al. 2005) for tuberculosis and lung disease together with an accredited training course in 2006. Validation studies across different settings and patient populations attested the system moderate to substantial values of inter- and intra-reader agreement (Den Boon et al. 2005, Dawson et al. 2010, Agizew et al. 2010, Hoog et al. 2011a, Pinto et al. 2013). A TB CXR Image Reference Set (TIRS) containing 17 A4 paper prints in a booklet was piloted as a practical, low-cost intervention to improve non-expert reading performance in a study in Malawi in 2010 and increased the number of correct decisions to initiate treatment modestly, but significantly (Waitt et al. 2013).

Interpreting chest radiographs is complex and subjective: it is a two dimensional representation of a three-dimensional structure, and there are varied manifestations of PTB. Widely used reading categories like abnormalities “consistent with TB” lack documented consensus on their exact definition and impair the transparency and comparability of reading results. While structured reporting methods like the CRRS allow an accurate itemization of the features seen on the radiograph, the reader's decision-making process at completion of the report remains concealed. Different scoring systems were developed to objectify this process and make it open to scrutiny (Ralph et al. 2010, Pinto et al. 2013, Breuninger, van den Hombergh, et al. 2014). Pinto et al. determined four CRRS features significantly associated with culture-positive tuberculosis and developed a weighted radiographic score (Pinto et al. 2013). In a cohort of 473 presumptive TB patients a score cut-off ≥ 2 improved specificity of the X-ray report considerably (63.9% vs. 27.5%) compared to the reader's subjective CRRS conclusion "consistent with active TB" without significant loss of sensitivity (85.5% vs. 93.4%) (Pinto et al. 2013).

At the same time, the Tanzanian Chest X-ray Initiative, a working group formed by implementing partners[*], radiologists and the TB/HIV collaboration Telemedicine Department of the Ministry of Health, Tanzania, compiled the Tanzanian Chest X-ray Score (TXS) (Appendix B). Aim of the TXS is to promote standardised, objective and reproducible CXR reporting and interpretation by accurate tabulation of radiographic features, assignment of attributable scores and translation of the cumulative score into conclusion categories. Depending on the combination of radiological features determined by the reader, the TXS labels the radiograph as 1.normal, 2.abnormal, not suggestive for active TB, 3.abnormal, consistent with active TB or 4.abnormal, highly suggestive for active TB. Scores in the first working version of the TXS are based on expert opinion. Until now, the implementation of the TXS has been limited to its descriptive part. It assisted readers in the Gambian prevalence survey in 2011 as a structured reading methodology, but no results on its performance have been published so far. The interpretative part and validity of the score have yet to be tested.

1.3.4 Computer-aided diagnosis of PTB

In contrast, the automated reading of radiographs by computers is devoid of inter- and intra-observer variability and independent of trained reading personnel on site. Research in this field started fifty years ago. Early optimistic goals such as "fully automating the

[*] PharmAccess Foundation, Netherlands and International Center for AIDS Care and Treatment Programs (ICAP) at the Columbia University's Mailman School of Public Health, USA

chest exam” (Conners et al. 1982) are still far from being achieved. However, at least one application, the automatic detection of masses and micro-calcifications in mammograms, has been successfully integrated into clinical routine to support radiologists in their decision (Samulski et al. 2010).

A growing number of research groups are working on promoting computer-aided diagnosis (CAD) in PTB. Given the variety of the disease’s manifestation on radiographs, some researchers have focused on the detection of specific features like cavities (Xu et al. 2011), nodules (Leibstein & Nel 2011), pleural effusion (Maduskar, Hogeweg, et al. 2013) or a miliary pattern (Koeslag & de Jager 2001) while others are trying to analyse the whole image. Automated observation, analysis and interpretation of a CXR usually include pre-processing steps (to enhance contrast, reduce anatomical noise and detect lung boundaries) that are followed by texture and/or geometry feature computation and classification methods (Jaeger et al. 2013). The recently published review article by Jaeger and colleagues summarises the efforts in this research area and demonstrates the variety of research directions and mathematical models used for this task (Jaeger et al. 2013). The authors conclude that even though proposed CAD algorithms seem to perform reasonably well when tested individually, no fair comparison can be made without testing the systems on the same, preferably large and publicly available dataset of well-characterized patients (Jaeger et al. 2013). A welcome effort was the publication of two CXR datasets from a population screening in the U.S. and a Chinese outpatient clinic by the U.S. National Library of Medicine last year (Jaeger, Candemir, et al. 2014). Unfortunately, only clinical reading reports, but no microbiological results are available for these images (Jaeger, Candemir, et al. 2014). Earlier, Jaeger et al. have tested a CAD prototype system on these datasets and achieved an area (A_z) under the receiver operating characteristic (ROC) curve of 0.87 and 0.90, respectively (Jaeger, Karargyris, et al. 2014). The system, which combines a segmentation method, texture and shape features, is currently being tested in a population screening in Kenya, but no results have been published so far (Jaeger, Karargyris, et al. 2014, Antani 2015). Limited information is available on DigiportXCAD (MVIP, Germany), another CAD solution for the detection of PTB on chest radiographs, which is currently under evaluation in Bangladesh (MVIP Software+Consulting GmbH & Verhey, FIND 2015).

CAD4TB, the most advanced CAD solution for the detection of pulmonary tuberculosis on chest radiographs to date, was developed by the Diagnostic Image Analysis Group (DIAG) at Radboud University Medical Center, Nijmegen, The Netherlands. CAD4TB is a software framework that integrates the results of different sub-systems for shape, texture and symmetry analysis. Its output is an abnormality score between 0 and 100, where 0 describes a normal chest radiograph and 100 a chest radiograph that is highly suggestive for pulmonary TB. The first CAD4TB beta prototype underwent field tests in 2010 and has been developed since then.

Previous software versions showed promising results: CADx, a research prototype, reached a sensitivity of 95% at a specificity of 57% in a pre-selected set of CXRs of homeless people in London (A_z=0.86) (Hogeweg et al. 2011). The next version, CAD4TB v1.08, detected culture-positive tuberculosis as accurate as four clinical officers among 161 presumptive TB patients in Zambia (Maduskar, Muyoyeta, et al. 2013). The yet lower accuracy (A_z=0.73) might be attributable to a very high prevalence of HIV among the study participants (Maduskar, Muyoyeta, et al. 2013). In the first prospective study of CAD4TB v1.08, Muyoyeta et al. tested the software in 350 presumptive TB patients in the same setting. Bacteriological confirmation by either Xpert or fluorescent microscopy was used as reference standard and attested the software a performance of A_z=0.71 (Muyoyeta et al. 2014). In this study population, the very high proportion of smear-negative patients (85%) degraded Xpert to a clearly sub-optimal reference standard with possible negative effect on the result (Muyoyeta et al. 2014).

1.4 Summary

The existing methods for a bacteriological diagnosis of pulmonary tuberculosis are insufficient in accuracy or too resource demanding for a widespread implementation in countries with a high burden of TB. An efficient triage and screening test to preselect individuals at highest risk for the disease for confirmatory testing is urgently needed to shorten diagnostic delays and ultimately close the case detection gap. With its high sensitivity, rapidly available results and high throughput capacity, chest radiography offers major strengths for this task. However, the absence of reporting standards, high levels of inter-reader variability and a lack of trained readers in most high-burden settings are serious drawbacks.

The development of structured reporting methods and interpretation aids for human readers as well as automated read-out solutions without the need for trained personnel are important steps to overcome the limitations of chest radiography. If these efforts prove accurate, reliable and without need for extensive training or resources, chest radiography qualifies as powerful triage and screening test.

We conducted the first validation study to assess the diagnostic accuracy of the CAD4TB software v3.07 on a large set of well-characterized adult presumptive PTB patients from sub-Saharan Africa. We compared the performance of the automated reading with the results of human observers of different experience levels. We evaluated the performance of the Tanzanian X-ray Score in the same patient population, tested the agreement between subjective and TXS conclusion, as well as the score's influence on inter- and intra-reader agreement.

The results of the CAD4TB validation have been published under the title *Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa* (Breuninger, van Ginneken, et al. 2014) in the open access journal PLoS One. Preliminary results of the TXS validation were presented at the 45th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union) under the title *Tanzanian X-ray score for the detection of active pulmoary TB on chest radiographs: a comparison with subjective assessment* (Breuninger, van den Hombergh, et al. 2014).

2 Objectives

2.1 Diagnostic accuracy of CAD4TB

In the first part of our study, we determined the diagnostic accuracy of the automated reading software CAD4TB for the diagnosis of pulmonary tuberculosis among symptomatic patients presenting to health care facilities in rural Tanzania. We evaluated differences in performance between sputum smear-positive and negative as well as HIV-positive and negative individuals. Results of the automated interpretation were compared with human reading results.

2.2 Evaluation of the Tanzanian Chest X-ray Score (TXS)

In the second part of our study, we evaluated the diagnostic accuracy of the Tanzanian Chest X-ray Score in the same patient population. For this purpose, we compared the reading results of a structured CXR report with subjective conclusion to the continuous and categorical output of the Tanzanian Chest X-ray Score. Focusing on the agreement between conclusion categories, we investigated whether the TXS conclusion represents the readers' opinion and the score's influence on inter- and intra-reader agreement of the X-ray report.

3 Materials & Methods

3.1 Study Population

This validation study was done on chest radiographs of participants from two cohort studies (TB Cohort and TB CHILD study) which had been conducted at the TB Clinic of the Ifakara Health Institute (IHI) in Bagamoyo, Tanzania. Tanzania has a high burden of active TB: according to the first national Tuberculosis Prevalence Survey in 2013, the prevalence is 295 cases per 100,000 population (Leth 2013). Bagamoyo, a town of 35,000 inhabitants, is located on the coast, approximately 70 km from the commercial capital Daressalam.

Individuals presenting with clinical signs and symptoms suggestive of pulmonary TB to surrounding primary health care facilities were referred to the IHI TB Clinic. Patients who met the inclusion criteria and gave informed consent were consecutively enrolled into either the TB Cohort or TB CHILD study. Patients who received anti-TB treatment during the last year, were severely sick or did not reside within the study area were not included. Recruitment of patients started in September 2010 and was completed in March 2012. In both studies, the patients were followed up for 5 to 18 months. The main objective of the TB Cohort study was to generate a sound understanding of TB epidemiology in the Bagamoyo region. The TB CHILD study was conducted to assess performance characteristics of new TB diagnostics in adults and children. Findings from the TB Cohort and TB CHILD study other than the CAD4TB and TXS validation results have been published elsewhere (Portevin et al. 2014, Reither et al. 2014, 2015, Mhimbira et al. 2015, Petrone et al. 2015, Kroidl et al. 2015).

For both studies written informed consent had been obtained from all literate patients. In case of illiteracy, informed oral consent had been attested by an impartial witness and documented with the patient’s fingerprint according to ICH GCP guidelines as approved by the IHI Institutional Review Board and the Medical Research Coordinating Committee of the National Institute for Medical Research, Tanzania.

All adult patients from the TB Cohort and TB CHILD study were eligible for the CAD4TB and TXS validation study if they initially presented with persistent cough of 2 weeks or more and at least one of the following TB associated findings: haemoptysis, chest pain, fever, night sweats, constant fatigue, recent unexplained weight loss, loss of appetite, malaise or contact with a known TB case.

3.2 Specimen collection & Laboratory methods

At enrolment, the participants answered a detailed questionnaire about their medical history, underwent clinical examination, had a posterior-anterior chest radiograph taken and sputum and blood samples collected. All CXRs (resolution: 1760 x 2140 pixel) were taken with a Philips Cosmos BS radiography system, which operated combined with a Philips PCR System Eleva S processor.

Two sputum specimens, one ‘spot’ and one early morning, were routinely obtained and used for acid-fast bacilli (AFB) smear and culture examination. All samples were decontaminated using the standard NALC-NaOH method, inoculated on both solid (Löwenstein-Jensen, LJ) and liquid (Mycobacterium Growth Identification Tube, MGIT) media and incubated at 37°C. Smears were performed from the decontaminated pellet, followed with Ziehl-Neelsen (ZN) staining. All positive cultures were tested by ZN microscopy for the presence of AFB and *Mycobacterium tuberculosis* was confirmed by MPT64 antigen and/or molecular tests (Genotype MTBC, CM or AS; Hain Lifescience, Nehren). Interpretation of all microbiological tests was carried out blind to clinical information and radiological results. Voluntary HIV counselling and testing was offered to all participants. The laboratory work was carried out according to Good Clinical Laboratory Practice to guarantee objective standards, quality control and assurance.

3.3 Classification

All patients were classified by the study physicians (M.D., 1-3 years of clinical experience) in consultation with a senior physician (M.D., 20 years of clinical experience) into seven groups (table 2) according to all clinical and microbiological information available 5 months after enrolment. Allocation to the groups was initially not mutually exclusive. However, for the purpose of this analysis, it was agreed that classification to either group A (s+/c+ M.tb) or B (s-/c+ M.tb) supersedes classification to C (s±/c+ NTM) or E (EPTB), and classification to either group G (Indeterminate) or D (s-/c- clin.TB) supersedes classification to group C (s±/c+ NTM). Patients with resolved symptoms after 5 months and who were confirmed to be definitely free of TB (group F) are referred to as 'Controls' in the following. An additional consistency check, which was done after the publication of the results (Breuninger, van Ginneken, et al. 2014), revealed initial misclassification for 10 of 861 patients. All statistical analyses were repeated and led to the same or marginally different results, which are reported in this thesis.

Group	**Description**	**Short form**
A	Smear positive/ culture positive, *Mycobacterium tuberculosis*	s+/c+ M.tb
B	Smear negative/ culture positive, *Mycobacterium tuberculosis*	s-/c+ M.tb
C	Smear negative or positive/ culture positive, nontuberculous mycobacteria (NTM), irrespective of clinical relevance	s±/c+ NTM
D	All cultures negative, CXR and clinical symptoms very suspect for PTB (clinically diagnosed TB)	s-/c- clin.TB
E	Cytologically/ histologically/ microbiologically confirmed extrapulmonary TB	EPTB
F	All smears and cultures negative and sustained recovery up to 5 months (e.g. resolved bronchitis or pneumonia)	Controls
G	Loss to follow-up after recruitment or any other combination of results (e.g. still symptomatic after 5 months)	Indeterminate

Table 2 Classification of the study population according to clinical and microbiological data

3.4 Reading of the chest radiographs

The readings of chest radiographs were carried out retrospectively for both, the automated and the human interpretation, and had no influence on the diagnosis of the study participants.

3.4.1 Automated reading

The computer-aided analysis of the CXRs was performed by the DIAG, independently and blind to clinical information and other radiological results. The radiographs were processed with the CAD4TB software v3.07. CAD4TB is a software framework including various subsystems operating on a pixel and image level for the detection of textural and shape abnormalities as well as for symmetry and correlation analyses (Hogeweg et al. 2010) (figure 2). The system is based on supervised machine learning and classifies new images upon prior training with labelled data (Maduskar, Muyoyeta, et al. 2013).

In CAD4TB, the analysis is broken down to several computable steps (van Ginneken et al. 2011): First, radiographs are pre-processed to normalise image features like resolution and grey scale (figure 3). A quick quality check follows and dismisses not correctly acquired posterior-anterior CXRs. During segmentation, the next step, the software seeks the anatomical orientation of the radiograph by demarcating structures like lungs, clavicles and ribs. Then, the defined lung fields are analysed for their local texture, shape and global symmetry (figures 4-6). In addition to that, a global correlation with a typical normal CXR is determined. Scores generated by these subsystems are combined to an overall score for each radiograph, which summarises the result of the automated analysis as an abnormality score for the presence of active disease between 0 - 100.

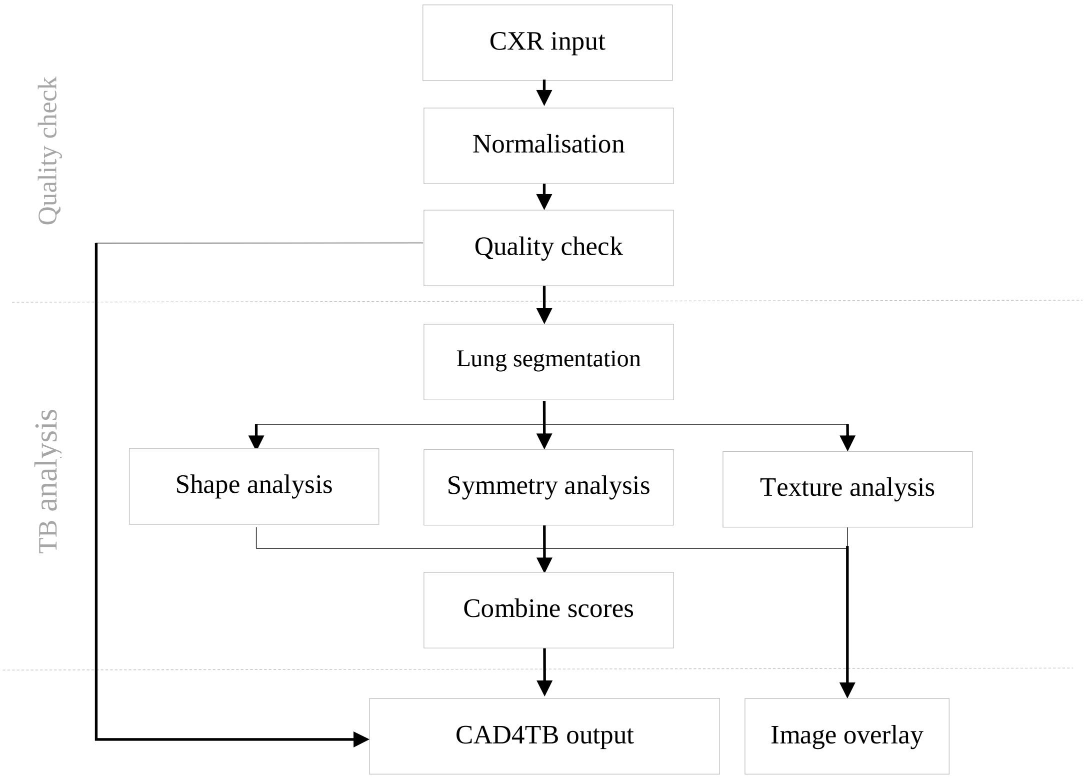


Figure 2 CAD4TB work flow.
Figure modified after Rick Philipsen et al., DIAG.

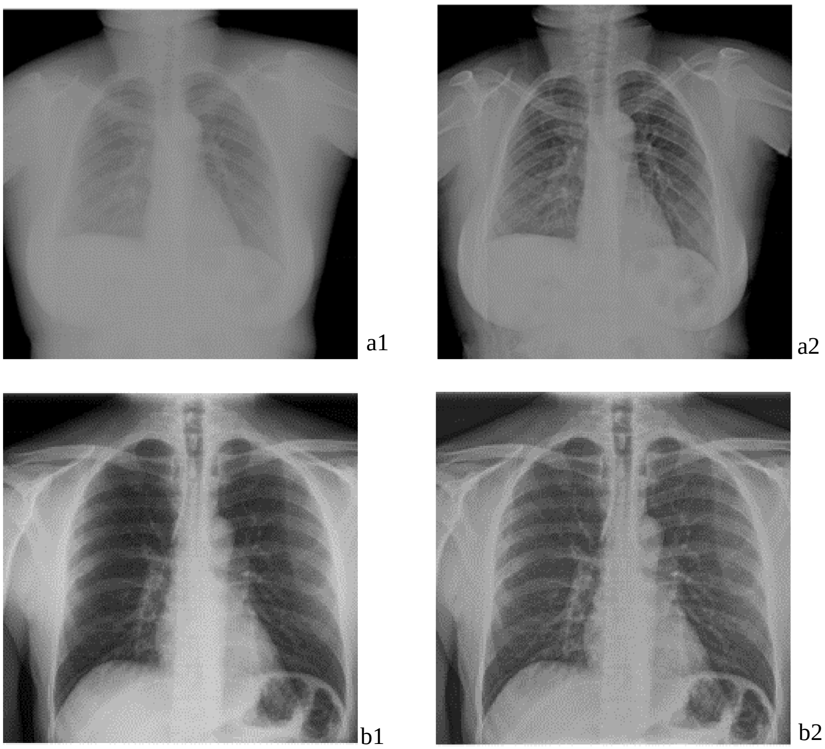


Figure 3 Normalisation
Images are pre-processed to normalise image features like resolution and grey scale (Philipsen et al. 2013). Images provided by Rick Philipsen from DIAG.

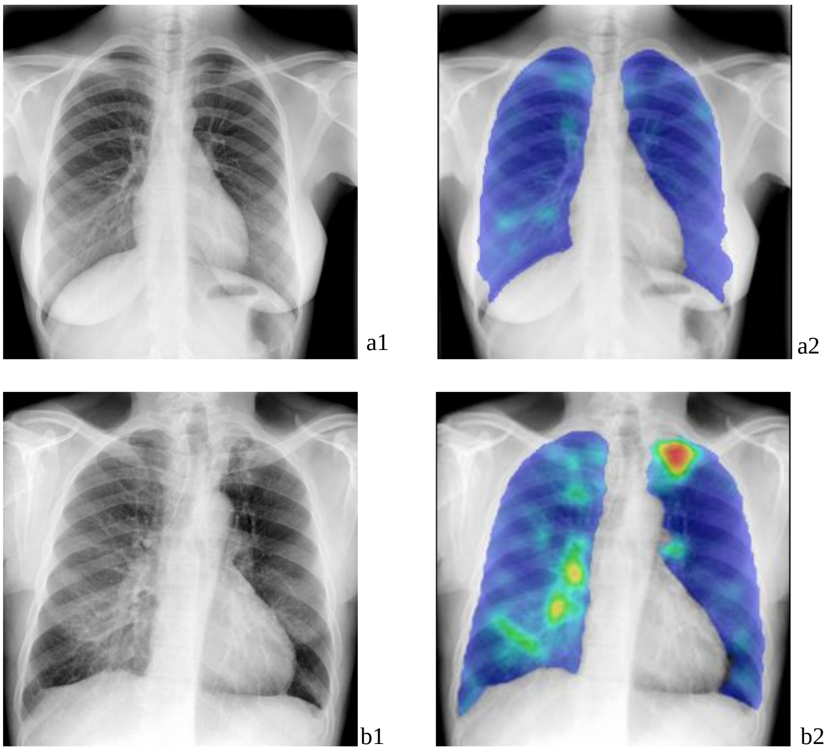


Figure 4 Texture analysis

Small circular patches are extracted, classified and labeled with an abnormality score. Patch labels are integrated to a textural abnormality score. The colour overlay depicts the degree of textural abnormality: blue = low textural abnormality, red = high textural abnormality. Images provided by Rick Philipsen from DIAG.

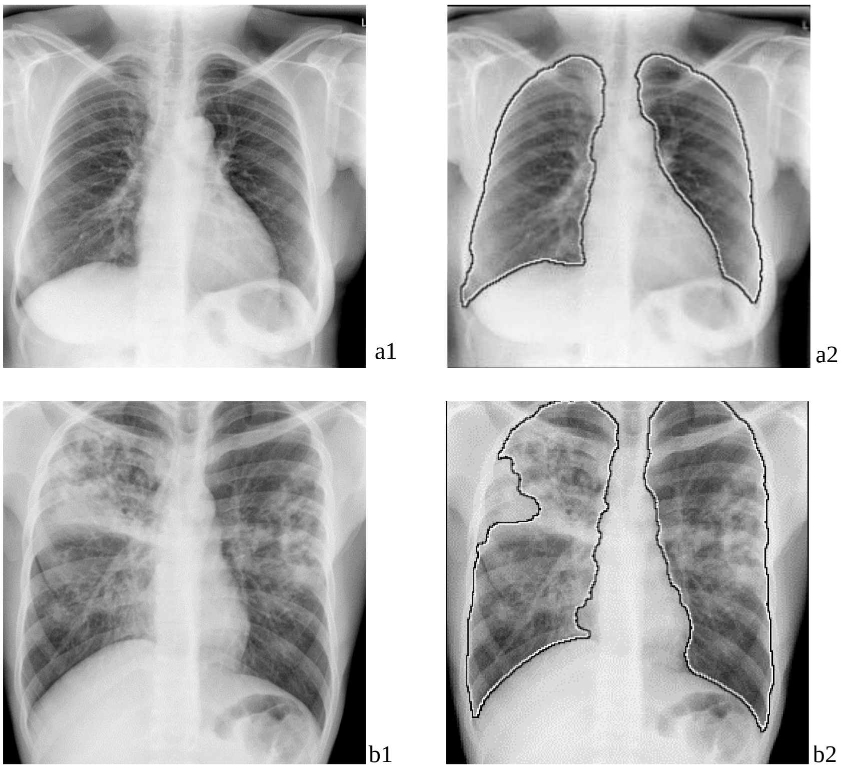


Figure 5 Shape analysis

Texture analysis can only consider abnormalities inside the outlined lung fields. The automated lung segmentation can be erroneous, especially if large abnormalities close to the pleural wall are present. An abnormal score of the shape analysis still reflects their presence (Hogeweg et al. 2010). Images provided by Rick Philipsen from DIAG.

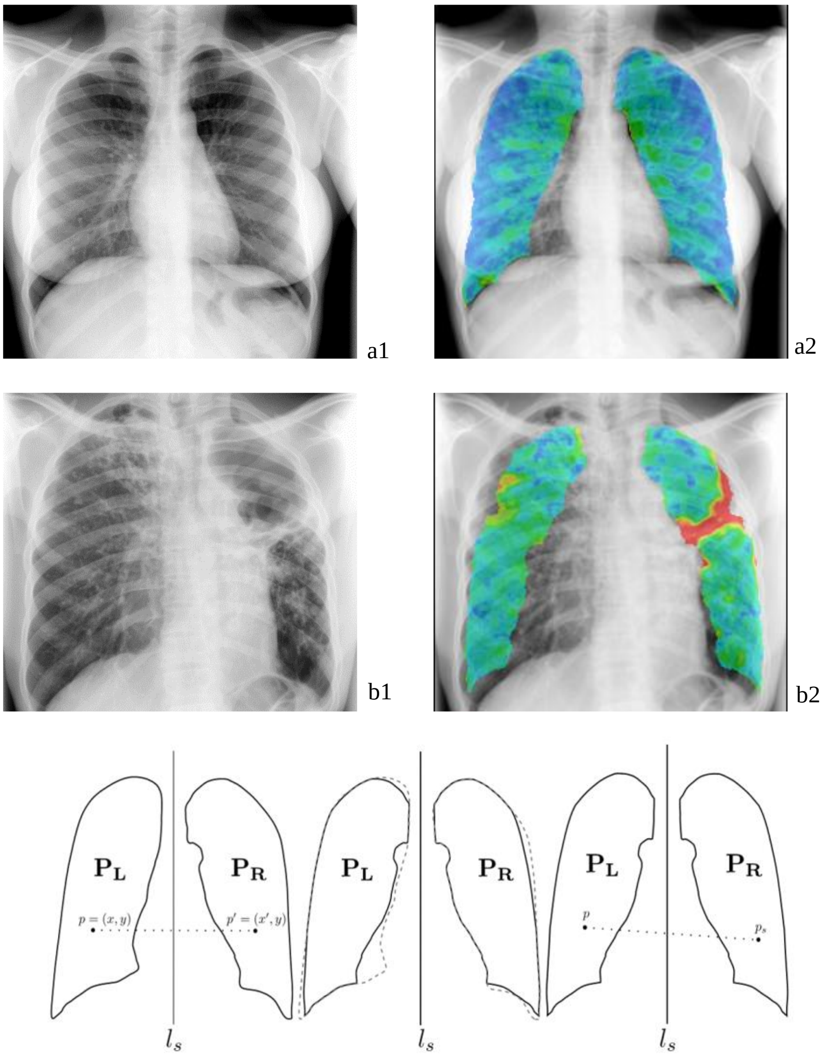


Figure 6 Symmetry analysis

The colour overlay depicts the degree of local symmetry: blue = high local symmetry, red = low local symmetry. Local symmetry for a point p is computed in a mirror symmetric set of locations (dashed vs. continuous lines in the middle sketch). The minimal dissimilarity of position and image characteristics between p and all points in P_R determines the optimal matching point p_s (Hogeweg 2013). Images provided by Rick Philipsen from DIAG.

3.4.2 Tanzanian Chest X-ray Score (TXS)

The TXS is a structured reporting and scoring system developed by the Tanzanian Chest X-ray Initiative to assist readers in the evaluation of chest radiographs for the presence of active pulmonary tuberculosis. It provides a computerised form with a list of radiographic findings and their location. Radiographic features are grouped into parenchymal, pleural, mediastinal and other abnormalities. A pre-defined score that - according to expert opinion - correlates to the grade of suspicion of active pulmonary TB disease was assigned to each feature (table 3).

Feature	Score
Parenchyma	
Consolidation	2
Cavitation	5
Isolated nodule(s) or tumour(s)	1
Multiple nodules or patchy infiltrate	3
Miliary pattern	5
Interstitial changes (other than nodules or miliary)	2
Fibrotic changes	1
Cystic change(s)	0
Solitary calcified nodule(s) or fibrotic scar(s)	0
Atelectasis or collapse (segment/lobe)	1
Pleura	
Pleural effusion (minor, unilateral)	2
Pleural effusion (extensive,<1/3 of lung visible, unilateral)	5
Pleural thickening	1
Pleural calcification	0
Mediastinum	
Hilar/mediastinal adenopathy	5
Extensive pericardial effusion (suspected)	5
Other	
Vertebral collapse, para-vertebral mass	5
Musculoskeletal abnormality	0
Cardiovascular abnormality	0
Diaphragmatic abnormality	0

Table 3 Radiographic features and assigned scores of the TXS

The scores of radiological features detected by the reader are added up to a cumulative score. Hence, in the first place, the output of the TXS can be "-" (none of the TXS features seen on the radiograph) or a value between 0 and 43 (all of the TXS features seen on the radiograph). This is the continuous output of the TXS. To give clinical significance to it, the developers of the TXS set a code for its translation into four different conclusion categories (table 4). Three hierarchical reading thresholds could be derived from these four conclusion categories ranging from considering only 'abnormalities highly suggestive for TB' (conclusion 4) over 'TB consistent abnormalities' (conclusion 3+4) to 'any abnormality' (conclusion 2-4) (table 4). This will referred to as the TXS conclusion of the X-ray report.

cumulative score	conclusion category	reading threshold for a positive test result		
- [*]	1. normal			
0-2	2. abnormal, findings not suggestive for active TB (TB sequel possible)	any abnormality		
3-4	3. abnormal, findings consistent with active TB, but TB sequel or other lung pathology possible		TB consistent abnormalities	
>5	4. abnormal, findings highly suggestive for active TB			abnormalities highly suggestive for TB
continuous TXS output	TXS conclusion			

[*] none of the TXS features seen on the radiograph

Table 4 TXS conclusion categories and reading thresholds

3.4.3 Human reading

In addition to automated reading with CAD4TB, three human observers of different experience levels read the same set of radiographs. Two of them were expert readers (one experienced chest physician, one radiologist) and one was a clinical officer with practical experience in reading chest X-ray exams in his role as District Tuberculosis and Leprosy Coordinator and a completed one week course on “X-ray interpretation of tuberculosis and HIV-related opportunistic infections among people living with HIV” (Tanzanian Ministry of Health and Social Welfare et al. 2010).

Each reader interpreted the whole set of radiographs once using the Tanzanian Chest X-ray Score (TXS, Appendix A) as a structured reporting template. First, they were requested to tick off the TXS chart according to their findings on the radiograph while being blind to the corresponding score values and score conclusion. After completion of the TXS chart, the readers were asked to diagnose each radiograph based on their own opinion and choose from one of the four conclusion categories (table 4). Consequently, each reader produced two possibly different results: one ‘TXS conclusion’ and one so-called ‘manual conclusion’. The readers were aware of the study’s inclusion criteria and the patients’ age but blind to clinical information, bacteriological results and the results of their co-readers.

For assessment of intra-reader variability, two of the three readers (one expert reader and the clinical officer) re-read the same random subset of 199 radiographs after a period of 4 weeks. As before, they reached a TXS conclusion and a manual conclusion. Again, they were blind to clinical information, bacteriological results as well as to their own and the other’s reading results from past or present.

3.4.4 Summary

In the first part of this study, the validation of CAD4TB, we compared automated reading results to the manual conclusion of all three human readers. To obtain the manual conclusion, the TXS was used without the score function, but as mere structured reporting form. In the second part of our study, we validated the scoring function of the TXS. For this purpose, we compared the continuous and the categorical score output (TXS conclusion) with both the manual conclusion and culture-confirmed M.tb.

3.5 Data analysis

Culture-confirmed M.tb was used as the reference standard to assess the diagnostic accuracy of CAD4TB and the human readers for the diagnosis of PTB. Individuals whose state of disease could be definitely determined were included in the analysis: group A (s+/c+ M.tb) and B (s-/c+ M.tb) as true cases and group F (Controls) as definite non TB patients. Secondary performance analysis was carried out in which individuals of group C (s±/c+ NTM) and E (EPTB) were considered additionally to group F (Controls) to be most likely free of pulmonary TB. Individuals of group D (s-/c- clin.TB) were classified partly due to an abnormal CXR and were thus excluded from analysis.

Receiver operating characteristic (ROC) curves and their areas under the curve (A_z) were calculated based on the CAD4TB output and the continuous output of the TXS. Their 95% confidence intervals (CI) and p-values were computed using the De Long method (DeLong ER et al. 1988). To plot a ROC curve, the true positive rate (sensitivity) of each possible test result is plotted against the respective false positive rate (1-specificity). A test, not better than random guess, will result in a diagonal line from the bottom left to the top right of the plot, since true and false positive rates are equally high. The area under this ROC curve or *line of no-discrimination* is 0.5. The better the discriminatory power of a test, the closer the curve is to the top left corner and the more equals its area under the ROC curve to 1.

The performance of human readers was summarised by calculating sensitivities, specificities, positive and negative predictive values as well as diagnostic likelihood ratios and their 95% confidence intervals for reporting ‘abnormalities highly suggestive for TB’ (conclusion 4), ‘TB consistent abnormalities’ (conclusion 3+4) or ‘any abnormality’ (conclusion 2-4). This was done for both the manual and the TXS conclusion. The same performance measures were calculated for several exemplary cut-offs of the CAD4TB software.

Proportions in different groups were compared using the chi-squared test. McNemar’s test was applied to compare the specificity of CAD4TB and humans at assumed levels of sensitivity. Mann-Whitney-Wilcoxon test was used to compare the CAD4TB scores between different groups.

Inter- and intra-reader variability for the manual and TXS conclusion was described by calculating the weighted Kappa agreement (κ_w), overall percentage agreement (p_o) and proportions of specific agreement for each conclusion category (p_s). A certain amount of agreement can occur purely by chance, this is the expected agreement. Kappa agreement contrasts observed agreement with expected agreement and therefore is considered as chance-corrected measure of agreement (Cohen 1960). Linear equal-spacing weights were applied to account for ordered conclusion categories (table 5). Kappa coefficients ≤ 0 were interpreted as poor agreement, 0.00-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and 0.81-1.00 as almost perfect agreement (Landis & Koch 1977). The overall percentage agreement is the ratio of the total number of ratings of agreement to the total number of ratings of disagreement. The proportions of specific agreement describe the probability of agreement between two readers per conclusion category. In this way, the contribution of agreement / disagreement on different conclusion categories to the overall percentage and kappa agreement can be inferred.

The significance threshold was set at p=0.05.

All calculations were done using the statistical software R, version 3.2.1 (R Foundation for Statistical Computing, Vienna , Austria) (R Core Team 2015) together with the extension packages ‘pROC’ (Robin et al. 2011), ‘epiR’ (Stevenson et al. 2013), ‘ggplot2’ (Wickham 2009), ‘reshape2’ (Wickham 2007), ‘plotrix’ (Lemon 2006), ‘obs.agree’ (Henriques T, Antunes L 2013) and ‘vcd’ (Meyer et al. 2015).

reader 1	**reader 2**			
	1	2	3	4
1	1.00	0.67	0.33	0
2	0.67	1.00	0.67	0.33
3	0.33	0.67	1.00	0.67
4	0	0.33	0.67	1.00

Table 5 Linear weights for agreement on conclusion categories between readers

3.6 Ethical considerations

All three studies (TB Cohort, TB CHILD and the CAD4TB and TXS validation study) were approved by the IHI Institutional Review Board and the Medical Research Coordinating Committee of the National Institute for Medical Research. Clearance certificates are included in the appendix.

4 Results

4.1 Characteristics of the study population

A total of 894 patients were enrolled in the CAD4TB and TXS validation study. Thirty-three patients had to be excluded from analysis because of an incomplete enrolment visit, pregnancy or missing chest radiograph (figure 7).

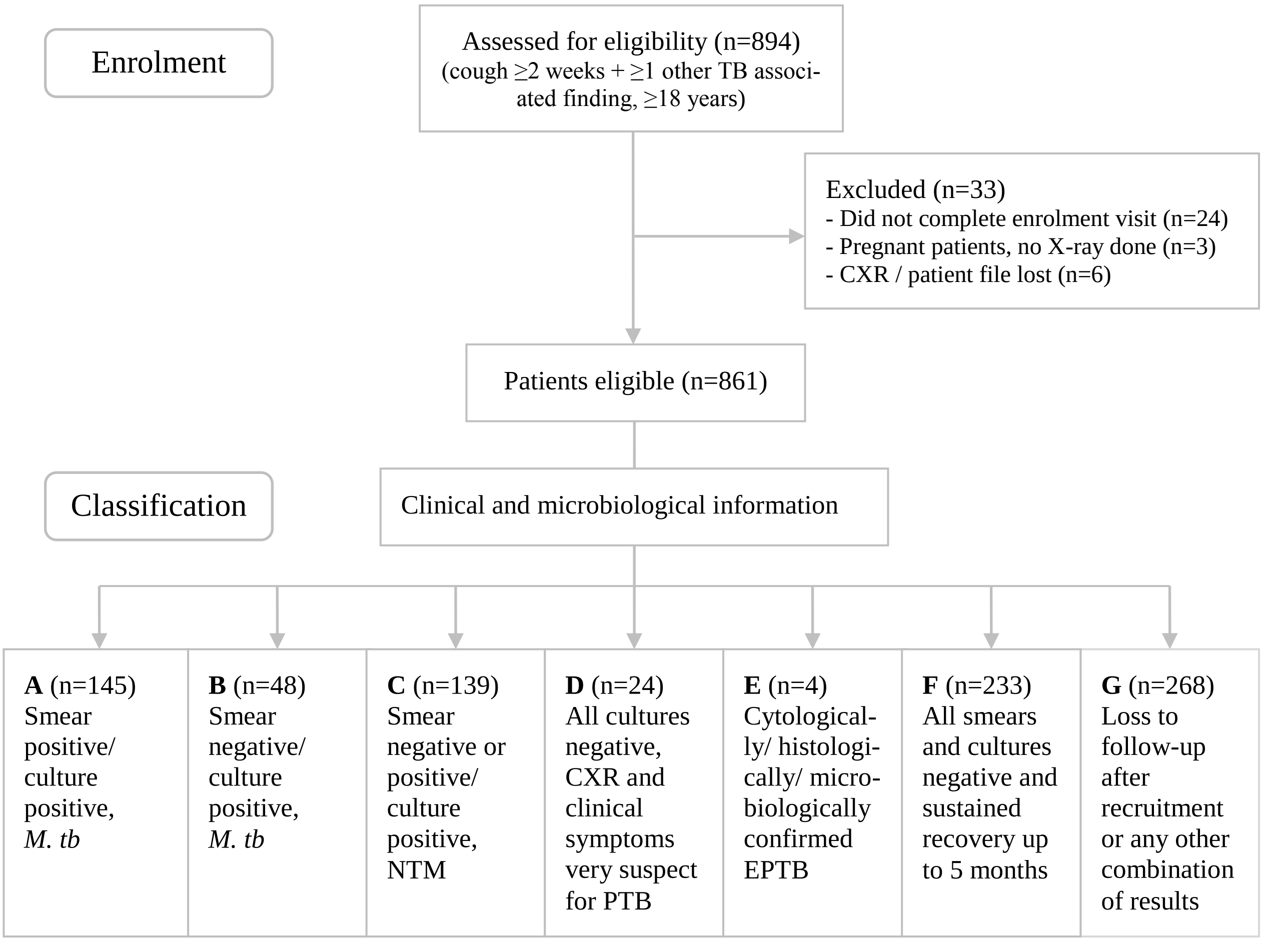

Enrolment
Assessed for eligibility (n=894)
(cough ≥2 weeks + ≥1 other TB associated finding, ≥18 years)
Excluded (n=33)
- Did not complete enrolment visit (n=24)
- Pregnant patients, no X-ray done (n=3)
- CXR / patient file lost (n=6)
Patients eligible (n=861)
Classification
Clinical and microbiological information
A (n=145) Smear positive/ culture positive, M. tb
B (n=48) Smear negative/ culture positive, M. tb
C (n=139) Smear negative or positive/ culture positive, NTM
D (n=24) All cultures negative, CXR and clinical symptoms very suspect for PTB
E (n=4) Cytologically/ histologically/ microbiologically confirmed EPTB
F (n=233) All smears and cultures negative and sustained recovery up to 5 months
G (n=268) Loss to follow-up after recruitment or any other combination of results

Figure 7 Flow chart of individuals taking part in the study

The final set of images for analysis consisted of 861 digital, posterior-anterior chest radiographs. Six of these radiographs were originally in a conventional film format and later digitised.

Group A (s+/c+ M.tb) and B (s-/c+ M.tb) included 193 (22%) of the study participants who were culture-positive for *Mycobacterium tuberculosis*. A further 233 patients (27%) presented with TB consistent symptoms but proved to be culture-negative with a sustained recovery after 5 months and were classified as group F (Controls) (figure 7).

Overall, the prevalence of HIV was 44%. There was a significant difference ($p<0.01$) between groups with the highest prevalence (73%, 95%CI 58-84%) in group B (s-/c+ M.tb) and the lowest (34%, 95%CI 26-42%) in group A (s+/c+ M.tb). The proportion of patients who reported a prior history of TB was 17% overall, but differed significantly ($p=0.02$) between classifications and was highest (50%, 95%CI 30-70%) among group D (s-/c- clin.TB) (table 6). Sex was evenly distributed (female sex =50%) in the study population as a whole, but with a significant difference ($p<0.01$) between classification groups and a far higher proportion of male (68%, 95% CI 59-75%) in group A (s+/c+ M.tb).

A pairwise comparison between culture-positive patients and controls revealed that culture-positive individuals (group A (s+/c+ M.tb) + B (s-/c+ M.tb)) were significantly more likely to suffer from night sweats (59 vs. 38%), fever (63 vs. 49%) and weight loss (68 vs. 44%) than individuals classified as group F (Controls) ($p\leq0.01$). There was no evidence of a difference in the frequency of haemoptysis between these groups ($p=0.34$) (table 6).

	No data	All	Group A s+/c+ M.tb	Group B s-/c+ M.tb	Group C s±/c+ NTM	Group D s-/c- clin.TB	Group E EPTB	Group F Controls	Group G Indeter- minate	p-value
Characteristic										
n (%)	0	861(100)	145(17)	48(6)	139(16)	24(3)	4(0)	233(27)	268(31)	NA
Mean age (standard deviation)	0	42(15)	37(13)	40(13)	42(15)	45(16)	38(11)	42(15)	44(16)	NA
Female sex n (%)	0	433(50)	47(32)	26(54)	78(56)	11(46)	4(100)	122(52)	145(54)	<0.01*
HIV-positive n (%)	4	379(44)	49(34)	35(73)	68(49)	10(42)	1(25)	92(39)	124(46)	<0.01*
History of TB n (%)	0	144(17)	17(12)	6(12)	27(19)	12(50)	0(0)	25(11)	57(21)	0.02*
Symptoms at first visit										
Cough ≥ 2 weeks n (%)	0	820(95)	138(95)	45(94)	134(96)	22(92)	3(75)	227(97)	251(94)	0.18
Night sweats n (%)	1	426(49)	91(63)	22(46)	68(49)	18(75)	3(75)	88(38)	136(51)	<0.01*
Haemoptysis n (%)	10	91(11)	11(8)	4(8)	16(12)	5(21)	0(0)	24(10)	31(12)	0.63
Fever n (%)	0	470(55)	91(63)	31(65)	80(58)	14(58)	3(75)	114(49)	137(51)	0.12
Weight loss n (%)	6	447(52)	100(69)	32(67)	64(46)	10(42)	3(75)	102(44)	136(51)	0.02*

Table 6 Summary statistics of study population

* significant differences between classification groups for a chi-squared test across all categories

4.2 Evaluation of CAD4TB

The distribution of CAD4TB scores (figure 8) for group A (s+/c+ M.tb) and D (s-/c- clin.TB) tends towards higher scores, this is less marked for group B (s-/c+ M.tb). The scores attained by individuals classified as group C (s±/c+ NTM) and F (Controls) are clustered around lower values but can be found across the whole range. Around one third of the individuals of group F (Controls) did attain a CAD4TB score greater than 50. On the whole, there is considerable overlap in the distribution of CAD4TB scores (table 7). The CAD4TB scores in group B (s-/c+ M.tb) are significantly lower than those of group A (s+/c+ M.tb) and higher than those of group F (Controls) (p<0.01).

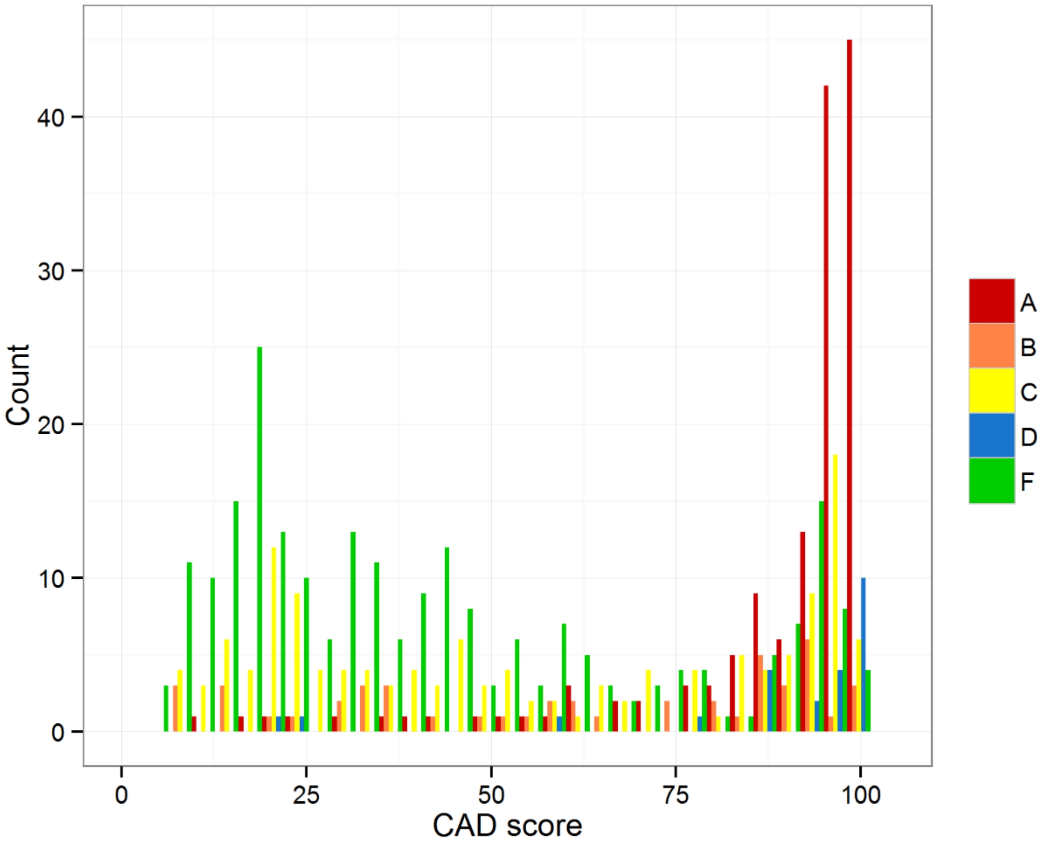


Figure 8 Distribution of CAD4TB scores

The x-axis denotes the CAD4TB score distributed between 0 and 100, the y-axis the number of patients with the respective CAD4TB score. Different colours represent different patient groups: group A (s+/c+ M.tb) = red, group B (s-/c+ M.tb) = orange, group C (s±/c+ NTM) = yellow, group D (s-/c- clin.TB) = blue and group F (Controls) = green.

		Median CAD4TB score	90% central range
All		61	11-100
Group A (s+/c+ M.tb)	all	97	43-100
	HIV+	92	32-100
	HIV-	97	60-100
Group B (s-/c+ M.tb)	all	63	11-99
	HIV+	62	13-98
	HIV-	86	12-96
Group C (s±/c+ NTM)		51	13-98
Group D (s-/c- clin.TB)		98	30-100
Group E (EPTB)		85	63-93
Group F (Controls)		34	9-94
Group G (Indeterminate)		57	11-100

Table 7 Median CAD4TB scores and 90% central range

4.2.1 Diagnostic accuracy of CAD4TB

The automated reading software was able to distinguish between culture positive PTB cases (group A (s+/c+ M.tb) + B (s-/c+ M.tb)) and non TB patients (group F (Controls)) with an area under the curve of 0.84 (95%CI 0.80-0.88). Including all M.tb culture-negative patients (group C (s±/c+ NTM), E (EPTB) and F (Controls)) as the negative reference standard, CAD4TB performed slightly, but not significantly, worse: A_z=0.81 (95%CI 0.77-0.85), p=0.26 (figure 9).

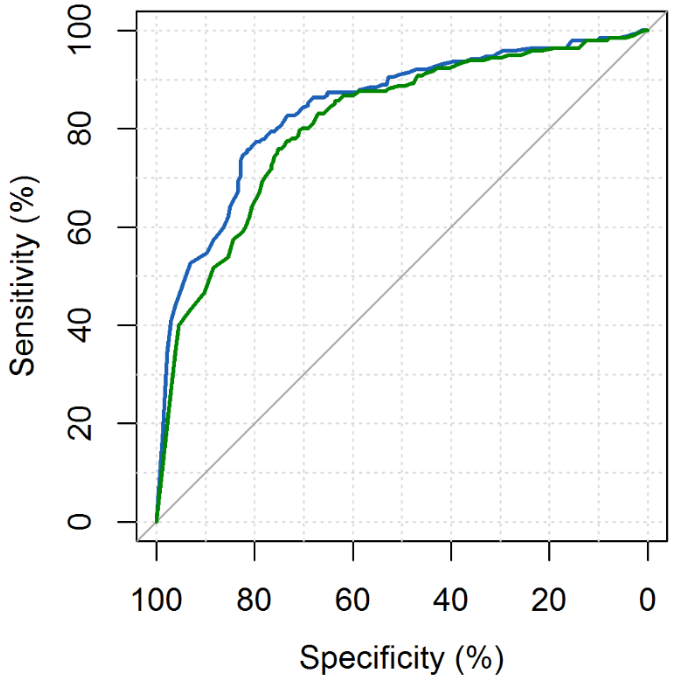


Figure 9 CAD4TB: ROC analysis for the detection of M.tb culture-positive individuals

— A (s+/c+ M.tb), B (s-/c+ M.tb) vs. F (Controls): A_z=0.84 (0.80-0.88),

— A (s+/c+ M.tb), B (s-/c+ M.tb) vs. C (s±/c+ NTM), E (EPTB), F (Controls): A_z=0.81 (0.77-0.85), p=0.26.

CAD4TB displayed a greater ability to differentiate smear-positive (group A (s+/c+ M.tb)) than smear-negative (group B (s-/c+ M.tb)) diseased individuals from non TB patients (group F (Controls)): A_z=0.90 (95%CI 0.87-0.93) against A_z=0.66 (95%CI 0.57-0.75), p<0.01 (figure 10).

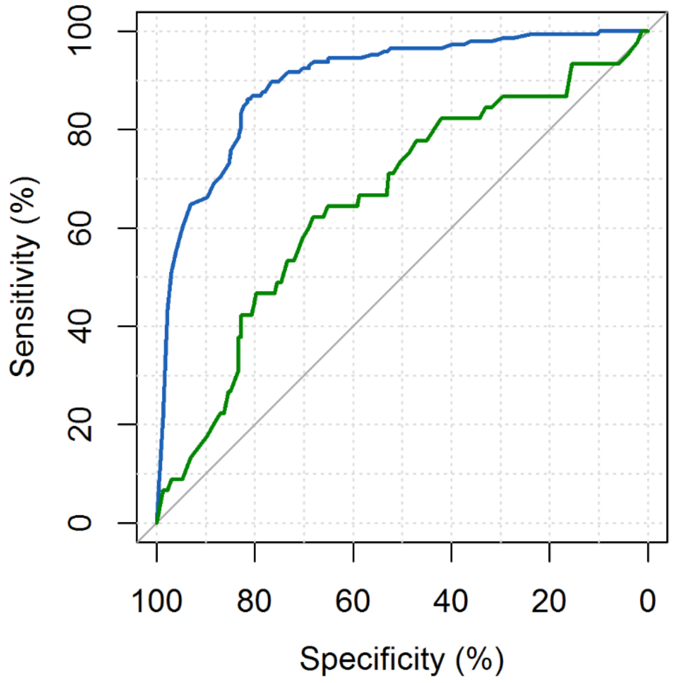


Figure 10 CAD4TB: ROC analysis for the detection of M.tb culture-positive individuals by smear status

🟦 A (s+/c+ M.tb) vs. F (Controls): A_z=0.90 (0.87-0.93),

🟩 B (s-/c+ M.tb) vs. F (Controls): A_z=0.66 (0.57-0.75), p<0.01.

Similarly, the software distinguished diseased individuals (group A (s+/c+ M.tb) + B (s-/c+ M.tb)) from non TB patients (group F (Controls)) significantly more accurately among the HIV-negative than among the HIV-positive patient population: A_z=0.89 (95%CI 0.85-0.94) against A_z=0.80 (95%CI 0.73-0.86), p=0.02 (figure 11).

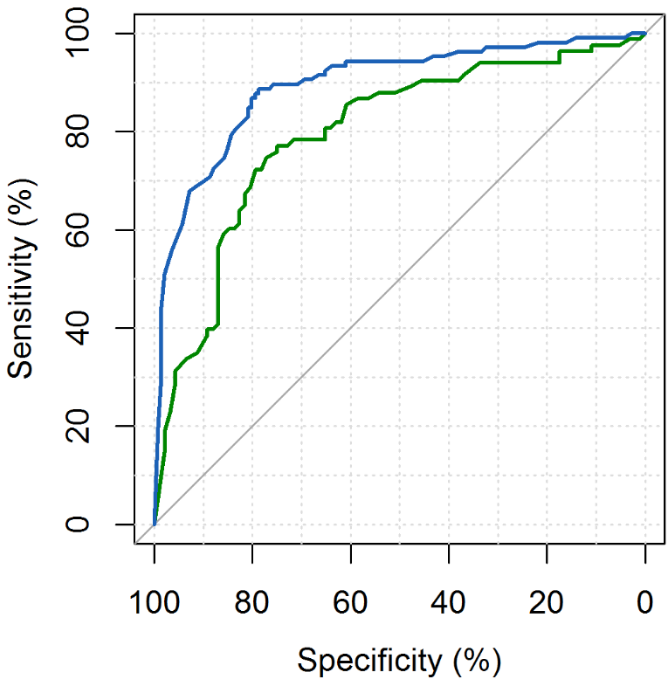


Figure 11 CAD4TB: ROC analysis for the detection of M.tb culture-positive individuals by HIV Status

—— HIV-neg. A (s+/c+ M.tb), B (s-/c+ M.tb) vs. F (Controls): A_z=0.89 (0.85-0.94),

—— HIV-pos. A (s+/c+ M.tb), B (s-/c+ M.tb) vs. F (Controls): A_z=0.80 (0.73-0.86), p=0.02.

Among group A (s+/c+ M.tb), B (s-/c+ M.tb) and F (Controls) there was no evidence of a difference in the performance of CAD4TB in between patients with and without history of TB: A_z=0.82 (95%CI 0.69-0.94) against A_z=0.84 (95%CI 0.80-0.89), p=0.67. The area under the curve of CAD4TB for the discrimination of group B (s-/c+ M.tb) against C (s±/c+ NTM) was 0.56 (95%CI 0.47-0.66).

We calculated a set of cut-offs of the CAD4TB score for our patient population (table 8). For example, a cut-off of ≥74 leads to a sensitivity and specificity of CAD4TB of 77% (95%CI 71-83%) and 80% (95%CI 74-85%), respectively. Optimal values of sensitivity cannot be obtained without a considerable trade-off of specificity, and vice versa.

	Threshold for test positivity	Sens.[1] [%] (95%CI)	Spec.[2] [%] (95%CI)	PPV[3] [%] (95%CI)	NPV[4] [%] (95%CI)	PLR[5] (95%CI)	NLR[6] (95%CI)
CAD4TB	≥23	95 (91-97)	33 (27-39)	54 (48-59)	89 (80-94)	1.41 (1.29-1.56)	0.16 (0.08-0.3)
	≥37	91 (85-94)	53 (46-59)	61 (55-67)	87 (81-92)	1.92 (1.66-2.21)	0.18 (0.11-0.28)
	≥56	85 (79-90)	69 (63-75)	69 (63-75)	85 (79-90)	2.74 (2.24-3.35)	0.22 (0.16-0.31)
	≥74	77 (71-83)	80 (74-85)	76 (69-82)	81 (76-86)	3.84 (2.94-5.01)	0.28 (0.22-0.37)
	≥89	62 (55-69)	85 (80-90)	78 (70-84)	73 (68-79)	4.26 (3.06-5.92)	0.44 (0.37-0.54)
	≥95	48 (41-55)	95 (91-97)	88 (81-94)	69 (64-74)	9.3 (5.26-16.45)	0.55 (0.48-0.63)

Table 8 Diagnostic performance of CAD4TB

[1] sensitivity, [2] specificity, [3] positive predictive value, [4] negative predictive value, [5] positive likelihood ratio, [6] negative likelihood ratio. All parameters were assessed against group A and B as positive reference standard and group F as controls.

4.2.2 Comparison of CAD4TB with manual conclusion of human readers

Setting the CAD4TB cut-off to give sensitivity values achieved by the human readers' manual conclusion allowed us to compare the performance of automated and human readings (figure 12, table 9). There was no evidence of a difference between the specificities achieved by the software and all three human readers reporting 'any abnormality' (p=0.67, 0.18, 0.49). This was different for tuberculosis specific reporting thresholds: CAD4TB was significantly more specific than the clinical officer was, but did not reach the accuracy level of the expert readers (p≤0.03).

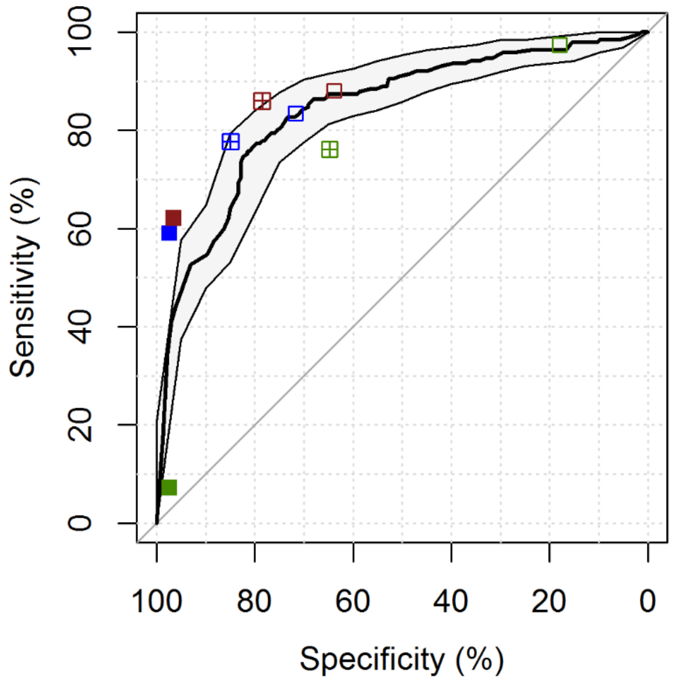


Figure 12 Comparison of automated and human reading

Legend. Sensitivity and specificity to distinguish group A (s+/c+ M.tb) and B (s-/c+ M.tb) vs. F (Controls). Line and shaded area: ROC curve and 95% CI for CAD4TB. The different colour of symbols represents different human readers: expert reader 1 = blue, expert reader 2 = red and clinical officer = green. The different fill of the symbols indicates different reading thresholds: empty symbols = 'any abnormality', crossed symbols = 'TB consistent abnormalities' and filled symbols = 'abnormalities highly suggestive for TB'.

human reading	expert reader 1	expert reader 2	clinical officer	
	□	□	□	any abnormality
	⊞	⊞	⊞	TB consistent abnormalities
	■	■	■	abnormalities highly suggestive for TB

A fourth reader, a senior radiologist with extensive experience in TB, reviewed the radiographs that had been rated false negative by CAD4TB at the exemplary cut-off (< 74) but as true positive (conclusion 3+4) by all human readers. This did not reveal any obvious pattern of abnormalities missed by CAD4TB.

4.3 Evaluation of the Tanzanian Chest X-ray Score

4.3.1 Diagnostic accuracy of the TXS

Comparing the areas under the ROC curves as performance measure for the diagnostic accuracy of the TXS revealed significant differences between the readers. Expert reader 2 rated the radiographs slightly more accurate than expert reader 1: A_z=0.88 (95%CI 0.84-0.91) against A_z=0.85 (95%CI 0.81-0.89), p=0.049. Both experienced readers outperformed the clinical officer clearly (A_z=0.67 (95%CI 0.61-0.72), p<0.001, figure 13). In other words, the probability of a randomly selected patient with PTB (group A (s+/c+ M.tb) and B (s-/c+ M.tb)) for receiving a higher TXS score than a randomly selected individual of group F (Controls) is 88% and 85% when his/her CXR is rated by one of the expert readers, but only 67% when interpreted by the clinical officer.

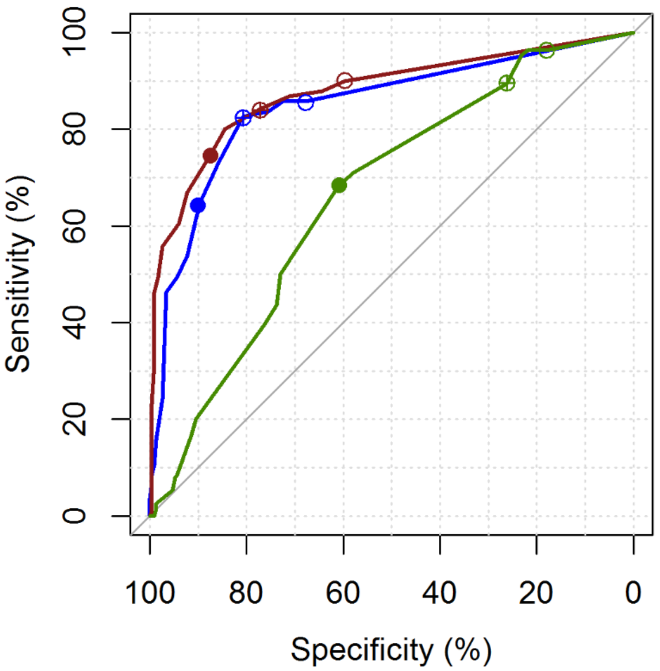


Figure 13 Diagnostic accuracy of human readers using the TXS

Legend. Sensitivity and specificity to distinguish group A (s+/c+ M.tb) and B (s-/c+ M.tb) vs. F (Controls) for expert reader 1 (blue), expert reader 2 (red) and the clinical officer (green) using the TXS. Lines denote the respective ROC curves for the continuous TXS output, while the circles denote the TXS conclusions ‘any abnormality’ (empty circles), ‘TB consistent abnormalities’ (crossed circles), or ‘abnormalities highly suggestive for TB’ (filled circles). The A_z of the experts’ ROC curves were significantly different from the one achieved by the clinical officer (p<0.001).

TXS continuous (= ROC curve)	—	expert reader 1	—	expert reader 2	—	clinical officer
TXS conclusion	○○○	any abnormality	⊕⊕⊕	TB consistent abnormalities	●●●	abnormalities highly suggestive for TB

4.3.2 Comparison of TXS with manual conclusion of human readers

The use of the TXS did not improve the readers’ ability to differentiate between culture-positive PTB cases (group A (s+/c+ M.tb) + B (s-/c+ M.tb)) and non TB patients (group F (Controls)), (figures 14.1-14.3: square symbols are found on or above the respective ROC curve). The clinical officer performed significantly better without the TXS when rating the CXRs for ‘TB consistent abnormalities’ as expert reader 1 did for ‘abnormalities highly suggestive for PTB’ (figure 14.1 and 14.3: the respective symbols are found above the 95% CI area). There was no evidence of a difference for all other reading thresholds.

Figure 14.1-3 Diagnostic accuracy of the TXS vs. manual conclusion of human readers

Legend. Sensitivity and specificity to distinguish group A (s+/c+ M.tb) and B (s-/c+ M.tb) vs. F (Controls) for expert reader 1 (blue), expert reader 2 (red) and the clinical officer (green) with and without the TXS. Lines and shaded area denote the respective ROC curves incl. 95% CI for the continuous TXS output. Different symbols denote different reading methodologies: circles for the TXS conclusions and square symbols for the readers' manual conclusion. The different fill of symbols indicates different reading thresholds: 'any abnormality' = empty symbols, 'TB consistent abnormalities' = crossed symbols, or 'abnormalities highly suggestive for TB' = filled circles.

TXS continuous (= ROC curve)	—	expert reader 1	—	expert reader 2	—	clinical officer
TXS conclusion	○○○	any abnormality	⊕⊕⊕	TB consistent abnormalities	●●●	abnormalities highly suggestive for TB
manual conclusion	□□□		⊞⊞⊞		■■■	

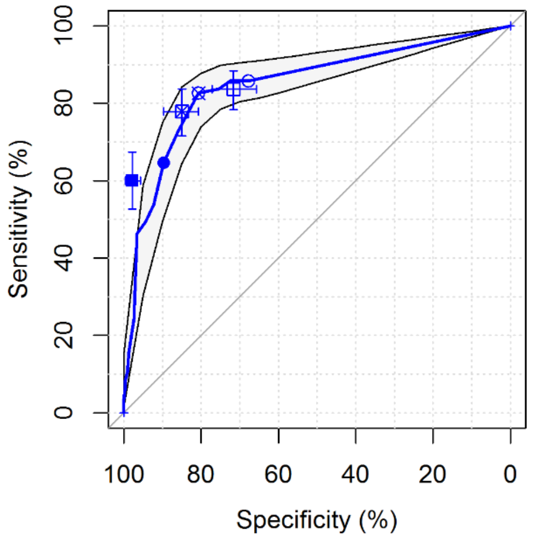


Figure 14.1 Diagnostic accuracy of expert reader 1.

TXS continuous output: A$_z$=0.85 (95%CI 0.81-0.89). Expert reader 1 detected abnormalities highly suggestive for TB significantly more accurate without the TXS.

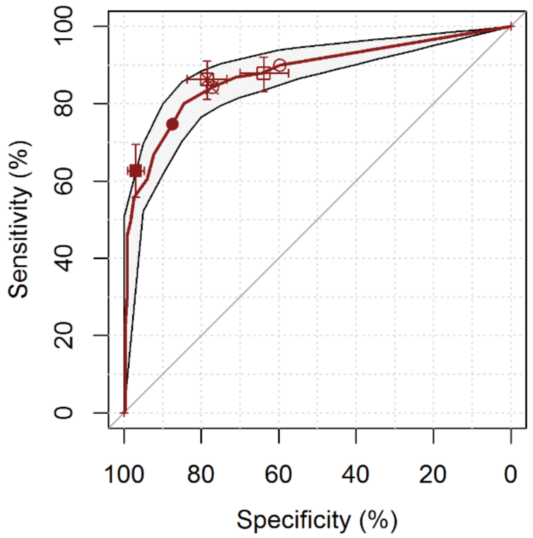


Figure 14.2 Diagnostic accuracy of expert reader 2.

TXS continuous output: A$_z$=0.88 (95%CI 0.84-0.91). There was no evidence of a difference between reading methodologies for expert reader 2.

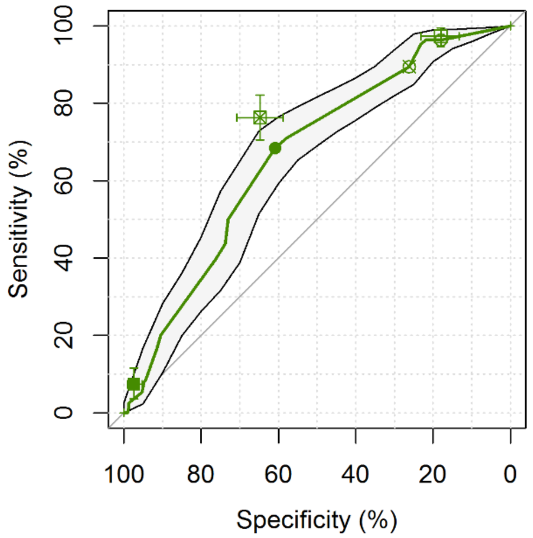


Figure 14.3 Diagnostic accuracy of the clinical officer.

TXS continuous output: A$_z$=0.67 (95%CI 0.61-0.72). The clinical officer detected abnormalities consistent with TB significantly more accurate without the TXS.

	reader	reading threshold	Sens.[1] [%] (95%CI)	Spec.[2] [%] (95%CI)	PPV[3] [%] (95%CI)	NPV[4] [%] (95%CI)	PLR[5] (95%CI)	NLR[6] (95%CI)
manual conclusion	**Expert reader 1**	highly suggestive for TB	59 (52-66)	97 (94-99)	95 (89-98)	74 (69-79)	22.94 (10.32-50.97)	0.42 (0.35-0.5)
		TB consistent	78 (71-83)	85 (80-89)	81 (75-86)	82 (77-87)	5.17 (3.78-7.09)	0.26 (0.2-0.34)
		any abnormality	83 (77-88)	72 (65-77)	71 (65-77)	84 (78-89)	2.94 (2.38-3.65)	0.23 (0.17-0.32)
	Expert reader 2	highly suggestive for TB	62 (55-69)	97 (93-99)	94 (88-97)	76 (70-80)	18.11 (9.08-36.1)	0.39 (0.33-0.47)
		TB consistent	86 (80-91)	79 (73-84)	77 (71-82)	87 (82-91)	4.01 (3.11-5.16)	0.18 (0.12-0.25)
		any abnormality	88 (83-92)	64 (57-70)	67 (61-73)	87 (81-91)	2.44 (2.04-2.92)	0.19 (0.13-0.28)
	Clinical officer	highly suggestive for TB	7 (4-12)	97 (94-99)	70 (46-88)	56 (51-61)	2.82 (1.1-7.19)	0.95 (0.91-1)
		TB consistent	76 (70-82)	65 (58-71)	64 (58-70)	77 (70-82)	2.16 (1.79-2.62)	0.37 (0.28-0.48)
		any abnormality	97 (94-99)	18 (13-24)	50 (44-55)	89 (77-96)	1.19 (1.11-1.27)	0.14 (0.06-0.36)
TXS conclusion	**Expert reader 1**	highly suggestive for TB	64 (57-71)	90 (86-94)	84 (77-90)	75 (70-80)	6.51 (4.35-9.73)	0.4 (0.33-0.48)
		TB consistent	82 (76-87)	81 (75-86)	78 (72-83)	85 (79-89)	4.27 (3.25-5.59)	0.22 (0.16-0.3)
		any abnormality	85 (80-90)	68 (61-74)	69 (62-75)	85 (79-90)	2.66 (2.18-3.23)	0.21 (0.15-0.3)
	Expert reader 2	highly suggestive for TB	75 (68-81)	88 (83-92)	83 (77-88)	81 (75-85)	5.99 (4.22-8.51)	0.29 (0.23-0.37)
		TB consistent	84 (78-89)	77 (71-82)	75 (69-81)	85 (80-90)	3.69 (2.89-4.71)	0.21 (0.15-0.29)
		any abnormality	90 (85-94)	60 (53-66)	65 (59-71)	88 (82-93)	2.23 (1.9-2.63)	0.17 (0.11-0.26)
	Clinical officer	highly suggestive for TB	68 (61-75)	61 (54-67)	59 (52-66)	70 (63-76)	1.75 (1.45-2.11)	0.52 (0.41-0.65)
		TB consistent	90 (84-94)	26 (21-32)	50 (45-56)	75 (64-84)	1.21 (1.11-1.33)	0.4 (0.25-0.63)
		any abnormality	96 (93-99)	18 (13-24)	49 (44-55)	86 (73-94)	1.18 (1.1-1.26)	0.2 (0.09-0.44)

Table 9 Performance of human readers: structured review with manual conclusion vs. TXS conclusion

[1]sensitivity [2]specificity [3]positive predictive value [4]negative predictive value [5]positive likelihood ratio [6]negative likelihood ratio.

All parameters were assessed against group A and B as positive reference standard and group F as negative control.

4.3.3 Agreement between conclusion categories

Inter- and intra-reader agreement for the manual and TXS conclusion was described with Kappa coefficients. A possible, but arbitrary interpretation of Kappa coefficients is: $\kappa \leq 0$ = poor, 0.00-0.20 = slight, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, 0.81-1.00 = almost perfect agreement (Landis & Koch 1977).

4.3.3.1 Agreement between TXS conclusion and manual conclusion

Both expert readers reached substantial agreement between their TXS conclusion and their manual conclusion after a structured review of the radiographs: κ_w (95% CI) = 0.80 (0.78-0.83) and 0.79 (0.76-0.82). Agreement between conclusion categories of both reading methodologies was fair for the clinical officer: κ_w (95% CI) = 0.27 (0.23-0.31). The proportion of specific agreement between TXS conclusion and manual conclusion was highest for normal CXRs (conclusion category 1) for all readers (p_s=0.91-0.96), table 10. This means, if expert reader 1 chose category 1 as his manual conclusion, the probability that his TXS conclusion was also in category 1 was 96%.

agreement between manual and TXS conclusion	expert 1	expert 2	clinical officer
p_o	0.77	0.74	0.25
p_s for category 1, 2, 3, 4	0.96, 0.62, 0.46, 0.67	0.94, 0.59, 0.48, 0.71	0.91, 0.22, 0.17, 0.15
κ_w (95% CI)	0.80 (0.78-0.83)	0.79 (0.76-0.82)	0.27 (0.23-0.31)

Table 10 Agreement per reader between their manual and TXS conclusion
- presented in overall percentage agreement (p_o), proportions of specific agreement for each category (p_s) and weighted κ agreement (incl. 95% CI), n=861.

4.3.3.2 Influence of the TXS on inter-reader agreement

Inter-expert agreement was substantial for either reading methodology and did not improve using the TXS: κ_w (95% CI)= 0.67 (0.63-0.70) vs. 0.64 (0.60-0.68), p_o= 0.66 vs. 0.63. The probability of agreement between the experts was much higher for conclusion categories 1 and 4 (‘normal CXR’ and ‘abnormalities highly suggestive for PTB’) than for conclusion categories 2 and 3 (‘abnormalities not suggestive for active TB’ and ‘abnormalities consistent with active TB’).

Reading results showed only fair agreement between the clinical officer and either expert reader for a structured review of the radiograph with manual conclusion: κ_w (95% CI)= 0.24 (0.21-0.28) and 0.25 (0.22-0.29). A higher overall percentage agreement with the use of the TXS (manual: p_o= 0.27, 0.28 vs. TXS: p_o= 0.35, 0.36) was accompanied with a downward trend of the chance-corrected Kappa agreement in this observer constellation (manual: κ_w (95% CI)= 0.24 (0.21-0.28) and 0.25 (0.22-0.29) vs. TXS: κ_w (95% CI)= 0.20 (0.16-0.23) and 0.21 (0.17-0.25), table 11).

inter-reader agreement	expert1 - expert2	expert1 – clinical officer	expert2 – clinical officer
- manual conclusion			
p_o	0.66	0.27	0.28
p_s for category 1, 2, 3, 4	0.84, 0.33, 0.44, 0.74	0.32, 0.18, 0.34, 0.12	0.34, 0.16, 0.36, 0.16
κ_w (95% CI)	0.67 (0.63-0.70)	0.24 (0.21-0.28)	0.25 (0.22-0.29)
- TXS conclusion			
p_o	0.63	0.35	0.36
p_s for category 1, 2, 3, 4	0.81, 0.34, 0.24, 0.69	0.32, 0.14, 0.20, 0.51	0.35, 0.14, 0.20, 0.53
κ_w (95% CI)	0.64 (0.60-0.68)	0.20 (0.16-0.23)	0.21 (0.17-0.25)

Table 11 Inter-reader agreement for different reading methodologies and reader constellations

- presented in overall percentage agreement (p_o), proportions of specific agreement for each category (p_s) and weighted κ agreement (incl. 95% CI), n=861.

4.3.3.3 Influence of the TXS on intra-reader agreement

Intra-reader agreement for expert reader 1 and the clinical officer was determined after re-reading a subset of 199 radiographs. Expert reader 1 reached substantial intra-reader agreement with almost identical Kappa and overall percentage agreement parameters with either reading methodology. Intra-reader agreement was fair for the clinical officer when he rated the radiographs manually (κ_w (95% CI)= 0.35 (0.25-0.46)), but deteriorated significantly with the use of the TXS (κ_w (95% CI)= 0.20 (0.09-0.32)). The proportions of specific agreement revealed a bias towards category 2 and 3 in the clinical officer's choice of the manual conclusion, which was not present with the use of the TXS (table 12).

intra-reader agreement	expert1	clinical officer
- manual conclusion		
p_o	0.66	0.59
p_s for category 1, 2, 3, 4	0.84, 0.46, 0.46, 0.66	0.30, 0.62, 0.64, 0.00
κ_w (95% CI)	0.68 (0.61-0.75)	0.35 (0.25-0.46)
- TXS conclusion		
p_o	0.67	0.52
p_s for category 1, 2, 3, 4	0.81, 0.43, 0.44, 0.72	0.27, 0.14, 0.38, 0.66
κ_w (95% CI)	0.68 (0.61-0.76)	0.20 (0.09-0.32)

Table 12 Intra-reader agreement for one expert reader and the clinical officer using different reading methodologies

- presented in overall percentage agreement (p_o), proportions of specific agreement for each category (p_s) and weighted κ agreement (incl. 95% CI), n=199.

5 Discussion

5.1 Summary and interpretation of study results

5.1.1 CAD4TB

Automating the interpretation of a chest radiograph for the detection of active pulmonary tuberculosis leads to objective, reproducible results and a standardised way of reporting. One of the main findings of our study is that the automated reading software CAD4TB v3.07 achieved a good diagnostic accuracy (A_z=0.84 (95%CI 0.80-0.88)) on a large set of CXRs of presumptive TB patients from sub-Saharan Africa. The accuracy of CAD4TB was slightly, but not significantly, worse in our secondary analysis using a binary classification of patients (M.tb culture-positive vs. negative), which we included for a better comparability with other diagnostic accuracy studies.

In our study, performance of automated and human reading was comparable when the observers considered ‘any abnormality’, the common threshold in triage and screening situations to qualify for confirmatory tests. For more TB specific reading thresholds, however, the software outperformed the clinical officer significantly but did not reach the accuracy of the expert readers. The software identified a significantly higher proportion of smear-positive compared to smear-negative, culture-positive individuals - most likely because smear-negative PTB patients tend to have more discrete or atypical radiographic features, especially in combination with HIV infection (Siddiqi et al. 2003). This assumption is substantiated by the fact that CAD4TB detected PTB cases significantly more accurately among HIV-negative than HIV-positive individuals.

The identification of active PTB cases among symptomatic individuals with abnormal CXRs due to other pulmonary conditions (e.g. pneumonia) or sequelae of tuberculosis remains challenging for both human and automated readers. This fact manifests itself in low specificity values as a consequence of the considerable overlap in the distributions of CAD4TB scores per defined patient groups. Moreover, the far higher proportion of patients with a history of TB among group D (s-/c- clin.TB) coincides with very high CAD4TB scores for this group. Although, this can be explained by the classification of these patients due to chest radiograph findings in the first place, it also opens the debate on whether human and automated readers were able to sufficiently discriminate between active and past TB disease.

The relatively high number of patients that were found to be culture-positive for NTMs (16%, group C) is not uncommon in the sub-Saharan African context (Fourie et al. 1980, Buijtels et al. 2009, Aliyu et al. 2013). This is probably largely due to contamination of culture samples either at patient level or from the environment as only few patients suffered from a pathogenetic relevant NTM infection that fulfilled the diagnostic criteria for a Nontuberculous Mycobacterial Lung Disease according to the American Thoracic Society (Griffith et al. 2007). The inability of CAD4TB to differentiate between patients of group B (s-/c+ M.tb) and C (s±/c+ NTM) might be due to the heterogeneity of group C (s±/c+ NTM). There are different types of NTM pulmonary diseases (NTM-PD), e.g. nodular bronchiectatic NTM-PD or cavitary NTM-PD with substantial overlap between those. The latter type is radiologically very similar to pulmonary TB (Kim et al. 2014) and it is quite probable that CAD4TB was not able to discriminate these two pulmonary mycobacterial disease manifestations. However, it is tempting to speculate on the performance of an automatic diagnostic tool as CAD4TB that has been trained on CXRs or CT-scans of patients with nodular bronchiectatic NTM-PD.

5.1.2 Tanzanian Chest X-ray Score

The TXS did improve neither the accuracy nor the reproducibility of human reading results in our study. Both expert readers achieved comparably high accuracy levels with and without the TXS, except expert reader 1, who was significantly more accurate with his manual differentiation of ‘abnormalities highly suggestive for TB’. The clinical officer did not benefit from the use of the TXS. His performance with either reading methodology was only moderate and he could detect ‘abnormalities consistent with TB’ significantly more accurate without using the TXS.

Both expert readers agreed substantially (κ_w= 0.80/ 0.79) between their manual and their TXS conclusion, which supports the validity of the TXS. They were able to transform their visual impression of the CXR into an accurate tabulation of features as required by the TXS form. The clinical officer might have been overwhelmed with the precise itemisation as his manual conclusion was more accurate than and agreed only fairly with his TXS conclusion (κ_w= 0.27). The use of the TXS as mere structured reporting form together with a simple, categorical conclusion code yielded substantial levels of inter-reader agreement between experts. However, neither their inter- nor intra-reader agreement increased with the use of the score function. While the effect of the score function on chance-corrected agreement was minor for comparisons that included at least one expert reader, the intra-reader agreement of the clinical officer deteriorated significantly when he relied on the TXS conclusion. Hence, the TXS proved to be least useful for the type of reader for which it had originally been developed. Different levels of experience and a rather complex reading code have likely contributed to the disagreement between expert and non-expert readers.

5.2 Strengths and limitations

This is the first study to validate the CAD4TB software v3.07 and the Tanzanian Chest X-ray Score for the diagnosis of pulmonary tuberculosis on chest radiographs of presumptive PTB patients. For the first time, CAD4TB performance for various cut-offs was stratified by smear and HIV status. A major strength of our study is the adherence to standards for reporting of diagnostic accuracy (STARD) (Banoo et al. 2010) and guidelines for reporting reliability and agreement studies (GRRAS) (Kottner, Audige, et al. 2011). The well-characterized study population and the use of a robust reference standard further substantiate our results.

This evaluation of CAD4TB is independent from its developers of the DIAG. They run the software on the provided set of radiographs, but were blind to clinical and microbiological information and not involved in the statistical analysis.

Another strength of our study is the direct comparison of automated and human reading on the same set of images. However, the degree to which this comparison can be generalised is limited due to inter-reader variability in the interpretation of chest radiographs and our ability to include only one clinical officer. It might well be that other clinical officers would have outperformed the CAD4TB software in our study. A pre-reading meeting and/or a reference set of CXRs to establish consensus on reading categories between the three raters might have been beneficial and could have diminished inter-reader variabilities. Another limitation of our study is the fact that it was conducted in only one high burden country. A repetition of the study in different settings will be necessary to assess generalisability of the results.

HIV infection seems to influence the diagnostic accuracy of CAD4TB, so our findings cannot be readily generalised to populations that differ significantly in their HIV prevalence. A further constraint of the study is the high proportion (31%, group G) of patients who either could not be followed up sufficiently to comply with the precise classification criteria or who were still non-TB patients, but symptomatic after five months and therefore could not be classified as group F (Controls). However, since a heterogeneous patient group is concerned and the data can be assumed to be missing at random, it can be postulated that the study results were not substantially influenced.

Kappa values are the standard method to report agreement of tests with binary and categorical output (Graham et al. 2002, Kundel & Polansky 2003, van Cleeff et al. 2005, Den Boon et al. 2005, Zellweger et al. 2006, Kottner, Audigé, et al. 2011, Maduskar, Muyoyeta, et al. 2013, Pinto et al. 2013). Yet, their dependence on observed disease prevalence impedes a comparison between studies (Feinstein & Cicchetti 1990, Kundel & Polansky 2003). Therefore and as recommended, we reported other agreement coefficients like the overall percentage agreement and proportions of specific agreement (Kundel & Polansky 2003) and included all agreement tables (Kottner, Audige, et al. 2011) in the appendix.

5.3 Comparison with other studies

Maduskar et al. evaluated the performance of a previous CAD4TB version and compared it to both, clinical officers rating the radiograph between 0-100 and the binary decision of an expert reader (as radiological reference) for the presence of TB consistent abnormalities (Maduskar, Muyoyeta, et al. 2013). The high accuracy of CAD4TB (A_z=0.91) attained for the radiological reference (Maduskar, Muyoyeta, et al. 2013) is consistent with our finding that CAD4TB approaches values of sensitivity and specificity achieved by the expert readers. We decided to use hierarchical reading thresholds, as we believe that this reflects the common radiological practice in a setting like ours. The diagnostic accuracy of CAD4TB for the bacteriological reference was higher (v3.07, A_z=0.81) using the newer version in our study compared to previous CAD4TB versions used in the studies of Maduskar (v1.08, A_z=0.73) and Muyoyeta (v1.08, A_z=0.71). This could either suggest advancement in the development of the software, which would be especially encouraging as we evaluated its performance on images obtained from a different X-ray machine than the one it had originally been developed for. On the other hand, the lower accuracy levels reported by Maduskar and Muyoyeta could be attributable to the very high proportion of HIV-positive (Maduskar, Muyoyeta, et al. 2013) and smear-negative patients (Muyoyeta et al. 2014) in their study populations as this corresponds to the inferior performance of CAD4TB in HIV-positive and smear-negative patients in our study.

Meanwhile results of four other studies using CAD4TB v3.07 have been published (Khan et al. 2014, Zaidi et al. 2014, Philipsen, Sánchez, et al. 2015, Steiner et al. 2015). A smaller study among presumptive TB patients at health care facilities in Pakistan compared the performance of the software with reading results of one clinical officer and two radiologists (Khan et al. 2014). The human review revealed that a high proportion of the radiographs that scored higher than 80 with CAD4TB showed abnormalities not suggestive for TB (Khan et al. 2014). This is in line with a relevant number of patients of group F (figure 8 and table 7) who attained a false positive high CAD4TB score due to other pulmonary pathologies in our study.

In another study from Pakistan, Zaidi et al. proposed a prediction model for pulmonary TB derived by logistic regression of the CAD4TB score, demographics and symptoms and with TB detected by Xpert MTB/RIF as binary outcome variable (Zaidi et al. 2014). The final model, which combines the CAD4TB score with information on the presence of cough > 2 weeks, age and gender of the patient, achieved an area under the ROC curve of 0.87 (Zaidi et al. 2014). This suggests an additional yield in accuracy of the model as validation studies on the standalone performance of CAD4TB for the detection of TB (confirmed by either culture or Xpert MTB/RIF) have reported values between 0.71 and 0.86 so far (table 17). This is encouraging, but unfortunately neither the additional yield of the model compared to CAD4TB alone, nor the exact prediction model itself have been published yet (Zaidi et al. 2014) hampering its external validation.

The two other studies using CAD4TB v3.07 evaluated automated chest radiography as either triage (Philipsen, Sánchez, et al. 2015) or screening test (Steiner et al. 2015). CAD4TB proved as a valuable tool to identify individuals at highest risk for PTB among 388 presumptive TB patients at a health centre in South Africa (Philipsen, Sánchez, et al. 2015). A CAD4TB score ≥ 85 increased Xpert throughput from 45 to 113 per day, reduced costs per screened subject from \$13.09 to \$6.72 and costs per noti-

fied TB case from \$90.70 to \$54.34 with a compromise in sensitivity from 78.9% (Xpert alone) to 67.6% (Philipsen, Sánchez, et al. 2015).
Steiner et al. confirmed the suitability of automated chest radiography as a screening test for PTB among 511 predominantly asymptomatic prisoners (Steiner et al. 2015). CAD4TB v3.07 interpreted the images without calibration for the local X-ray equipment or need for additional intervention by the operators reliably and without diagnostic delay (>99%) (Steiner et al. 2015). The performance of the software was compared to a varied sample of readers at different levels of experience and a radiological reference determined by the consensus of two experienced TB radiologists (Steiner et al. 2015). The results are in line with prior comparisons of human and automated reading and our study findings as the software performed comparably to or better than a majority of less trained readers, but did not reach the accuracy of expert readers (Ginneken et al. 2012, Maduskar, Muyoyeta, et al. 2013, Breuninger, van Ginneken, et al. 2014, Philipsen, Sánchez, et al. 2015, Steiner et al. 2015). In comparison with substantial variation of human performance due to different experience levels (figure 17, Appendix F), CAD4TB performance seems to be highly consistent across different settings. A synoptic table of published results from CAD4TB validation studies can be found in the Appendix D (table 17).

As mentioned earlier, Pinto et al. were the first to derive and validate a solely radiographic score for the diagnosis of PTB (Pinto et al. 2013). To calculate this score, weights (in brackets) are assigned to four CRRS features: large upper lobe opacity (2), cavity (2), unilateral pleural effusion (1) and hilar or mediastinal lymphadenopathy (2) (Pinto et al. 2013). Figure 15 illustrates the score performance (triangle symbols and orange line) and compares it to the accuracy of a conventional CRRS report (orange square symbol) as well as the performance of the TXS (blue, red and green line) and CAD4TB (black line) in our study. Even though tested in different populations of presumptive PTB patients, this comparison reveals interesting aspects: When applied by experienced readers, Pinto's score with thresholds of ≥ 1 and ≥ 2 reached similar accuracy levels as CAD4TB and both our expert readers with the TXS, but with far less radiographic features considered. However, the TXS used by expert readers was considerably more accurate than Pinto's score at cut-offs ≥ 3. One of our main findings is that the clinical officer performed significantly worse, when he scored CXRs for TB consistent abnormalities with the TXS rather than relying on his manual conclusion. A possible reason for this might be the complexity of the TXS: it entails a wide range of possible radiographic abnormalities and less experienced readers might not be certain about their exact morphologic correlations. Non-expert readers, like the clinical officer in our study, might benefit from the training on and the use of a simpler score. The score proposed by Pinto et al., even though simple, yet needs to be tested by non-expert readers and in a population different from the one that it was derived in.

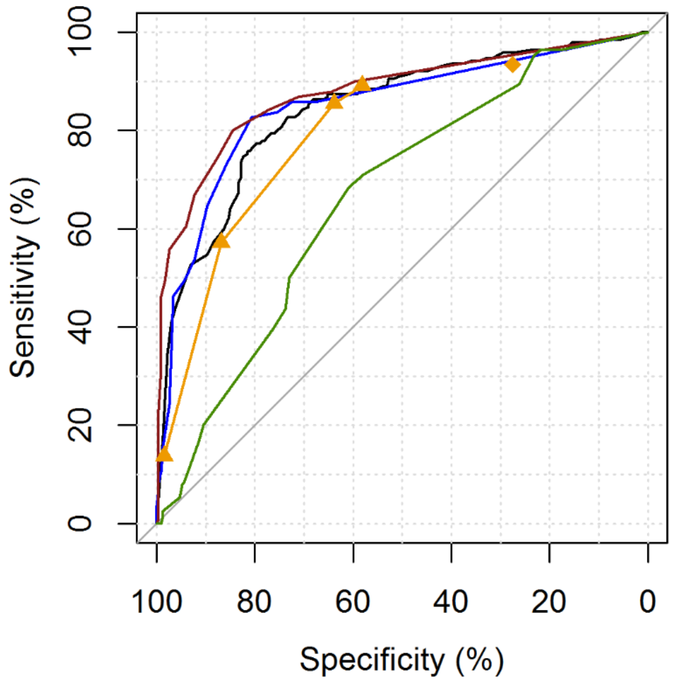


Figure 15 Comparison of the performance of TXS, CAD4TB and a score proposed by Pinto et al.

Legend. Diagnostic accuracy of the TXS and CAD4TB compared to the performance of a score proposed by Pinto et al. (evaluated in a different study population).

Pinto	—▲—	**TXS**	expert reader 1	—	**CAD4TB**	—
CRRS	◆		expert reader 2	—		
			clinical officer	—		

It is difficult to compare studies on inter- and intra-reader agreement of X-ray reports due to Kappa's dependence on observed prevalence (Feinstein & Cicchetti 1990, Kundel & Polansky 2003), differences in the number and experience of readers. Furthermore, due to the existing variety of reading methodologies – including structured reporting of features seen, categorical reading codes and grading the radiograph on a scale between 0-100 for the presence of active TB – such comparisons prove to be challenging. Agreement between both expert readers using the TXS as mere structured reporting form was substantial (κ_w = 0.67) and comparable to CRRS agreement levels (κ = 0.47-0.69) between experts on their conclusion of a TB consistent chest radiograph

(Den Boon et al. 2005, Dawson et al. 2010, Pinto et al. 2013). Together these study findings endorse the use of a structured reporting method. Intra-reader agreement levels with the CRRS were reported in only one study and were considerably higher than in our study (κ = 0.85-0.90 vs. κ_w = 0.68) (Den Boon et al. 2005). Only one other study has tested inter- and intra-reader agreement in non-expert readers and yielded much higher levels of agreement for clinical officers reporting 'any abnormality' in a Kenyan prevalence survey (table 17) (Hoog et al. 2011a). The substantial inter-expert agreement in our study may also be attributable to the use of a simple four-point categorical conclusion code. This is in line with previous studies that reported satisfactory agreement levels for other conclusion codes (Graham et al. 2002, van Cleeff et al. 2005, Zellweger et al. 2006, Shah et al. 2009). Although comparable, the diversity of PTB reading codes found in the literature is striking: each of five studies used an individual code (Graham et al. 2002, van Cleeff et al. 2005, Zellweger et al. 2006, Shah et al. 2009, Story et al. 2012).

A synoptic table of results from studies on the reproducibility and diagnostic accuracy of different CXR reading methodologies for the detection of TB is presented in the Appendix E (table 18).

5.4 Implications of findings & Outlook

Chest radiography can serve as either diagnostic, triage or screening test for PTB with varying requirements. The automated reading software CAD4TB and the structured reporting and scoring form TXS are two efforts to overcome its known drawbacks: inter-reader variability, the absence of reporting standards and lack of skilled readers in resource-constrained settings.

CAD4TB computes absolutely reproducible and standardised results without the need for trained reading personnel. Even though our findings attested the software a good diagnostic accuracy for culture-positive PTB among presumptive TB patients seeking care, its usefulness remains limited in this situation. Sputum smear microscopy is the primary method to diagnose TB in most high-burden settings (Denkinger et al. 2013, Kik et al. 2014). The current national diagnostic algorithm for presumptive adult TB patients in Tanzania, as in many sub-Saharan African countries, request between two to six negative sputum smear examinations and a failed treatment with a broad-spectrum antibiotic for 7 days before a chest X-ray is ordered (WHO 2011, NTLP Tanzania 2013). According to recommendations of the WHO, the CXR exam should precede an administration of antibiotics in settings where HIV is highly prevalent and resources are constrained (WHO 2007). CAD4TB specifically answers the question of an image's consistency with active PTB, yet at the expense of all other information the radiograph could offer to an experienced observer. Its output, a single number, does not reflect the presence of abnormalities unrelated to TB, whose detection might be not less important or even prompt immediate action (such as pneumonia, pneumothorax or lung tumours). Hence, it cannot replace a thorough X-ray report and its integration with clinical information by a medically trained person for the final diagnosis of smear-negative PTB when microbiological testing beyond sputum smear microscopy is not feasible.

In triage and screening situations, by contrast, the very condensed output of the automated reading might be preferable for the binary decision of either selecting a screened individual for confirmatory testing or declaring the absence of PTB. A strong feature of CAD4TB in this context is its continuous output, which allows adjusting the reading threshold to the purpose of use, local epidemiology and availability of resources (such as the capacity to perform culture or the number of Xpert MTB/RIF cartridges). Immediately available CAD4TB results enable a successful multi-step screening procedure.

There has been an extensive rollout of Gene Xpert MTB/RIF machines and test cartridges in 110 high-burden and low/middle-income countries since the WHO endorsed the test in 2010. The WHO strongly recommends its use as the initial diagnostic test in adults and children assumed to suffer from MDR-TB, HIV-associated TB or TB meningitis (WHO 2013b). Resource constraints make this recommendation conditional in other presumptive pulmonary and extrapulmonary TB patients (WHO 2013b). A cost-affordability analysis calculated costs to be five-fold higher for an indiscriminate use of Xpert as initial test in all patients with signs and symptoms of TB compared to the conventional diagnostic algorithm of smear microscopy and follow-on CXR, in case of smear-negativity (Pantoja et al. 2013).

In line with this, the WHO defined four high-priority target product profiles for new tuberculosis diagnostics in a consensus meeting in 2014 (WHO 2014c). Among these is a community-based triage or referral test for identifying people suspected of having TB for an efficient use of confirmatory tests (e.g. Xpert MTB/RIF) (WHO 2014c). The minimal requirements of this target product profile include a sensitivity > 90%, specificity > 70%, simple sample preparation, maintenance and calibration, at less than US$ 2.00 after scale-up and with a time to result of less than 30 minutes (WHO 2014c). In fact, assuming a widespread implementation of robust digital radiography equipment, CAD4TB only lacks a higher specificity to fulfil all of the above-mentioned criteria. In our study, CAD4TB achieved a specificity of 53% at a sensitivity of 91% for a threshold of ≥ 37. The software is under constant development and possibly subsequent versions will achieve the required accuracy. However, already at current accuracy levels, an overview table from the manufacturer of different CAD4TB thresholds with respective sensitivity and specificity values can guide the health worker in his decision of selecting a patient for confirmatory testing. Meanwhile, CAD4TB v4.10 has been released and received CE-certification for the use on any digital X-ray platform in 2015 (Boyle & Pai 2015). First results on its accuracy were presented at the Union World Conference on Lung Health in December 2015 and suggest further advancement of the software. CAD4TB v4.10 was tested retrospectively on 12,256 CXRs from DetecTB, an intensified case finding project in prisons and high-risk communities in Philippines (Philipsen, Sanchez, et al. 2015) and on 4,552 CXRs from the Gambian prevalence survey in 2011-2013 (Maduskar et al. 2015). In both studies the software achieved very high accuracy levels (A_z=0.91 and A_z=0.90) that were only slightly worse or identical to those of human readers (Maduskar et al. 2015, Philipsen, Sanchez, et al. 2015). Verification bias was present in both studies and hampers this comparison as only individuals with either positive symptom or CXR screen received confirmatory testing by either Xpert or sputum microscopy and culture examination (Maduskar et al. 2015, Philipsen, Sanchez, et al. 2015).

Non-availability of working X-ray equipment is a concern, but a growing number of manufactures focus on the development of digital radiology solutions for low- and middle income settings (FIND 2015). Radiography is essential for a wide range of diagnostic purposes and possible benefits will outrange its application in TB screening and diagnosis (Maru et al. 2010). The automated read-out solution CAD4TB presents a viable alternative to human readings of CXRs in screening and triage situations. Nonetheless, training of readers remains indispensable and must accompany the scale-up of digital radiography to prevent misuse, over- or under-diagnosis and initiation of inappropriate treatment. Quality assurance of CXR interpretation also warrants the use of reading standards.

The TXS was designed to assist less experienced readers in the interpretation of a CXR for the presence of PTB. However, the performance of the clinical officer, which was already moderate, deteriorated further with the use of the scoring function. This assessment is limited by relying on data derived from the reading of a single clinical officer. Nevertheless, in the absence of other evidence, we can not recommend the use of the TXS in its current form for non-expert readers. By contrast, substantial levels of inter-expert agreement let us endorse the use of the descriptive part of the TXS as reporting standard for expert readers. The expert readers themselves did not benefit from the scoring function of the TXS in terms of accuracy or reproducibility. Yet, high levels of agreement between their manual and TXS conclusion showed that the decision making process at the end of the X-ray report can be objectified and made transparent. Eventually, the TXS may assist less experienced readers not as reading, but as teaching tool. A reference set of radiographs with the respective TXS form completed by an expert reader leading to a conclusion open to scrutiny, together with culture results of the patient, can form a useful training tool. This should be accompanied by an international consensus on definition of terms and reading categories as research results cannot be compared without the use of comparable terms.

Further research, to test which score features can be omitted without compromise in accuracy, is in progress. Ideally, the result will be a simple, yet accurate score. Non-experts then could be specifically trained in the identification of these remaining features. Ultimately, an accurate and reliable score could assist non-expert readers in both, passive and active case finding situations. In settings where microbiological testing beyond sputum smear microscopy is not available, a relatively higher cut-off can support the diagnosis of smear-negative TB and the decision whether to initiate anti TB treatment. In triage and screening settings, a relatively lower score can be used to pre-select patients for confirmatory testing by culture or Xpert MTB/RIF.

The WHO set the goal to end the global TB epidemic by 2035 (WHO 2015d). This requires a much sharper decline in incidence than the average 1.5% per year during the last decade (WHO 2015d). In the absence of an effective vaccine, the only way to control the epidemic is to find and diagnose all TB patients, offer them effective treatment and render them non-infectious. However, with an estimated 3.6 million people living with undetected TB, who in turn can infect up to 10 other persons per year, we constantly lag behind the epidemic. Hence, key components of the END TB strategy are the early diagnosis of TB and the systematic screening of TB contacts and high-risk groups (WHO 2015d). A robust CAD has the potential to enhance and facilitate the implementation of these recommendations by ensuring high test standards of objectivity, reproducibility and accuracy in triage and screening without straining personnel resources. Automated reading solutions like CAD4TB are not yet able to rate a radiograph with the clinical reasoning of an experienced observer, who integrates the patient’s history, condition and characteristics with other diagnostic findings. Thus, for a more differentiated interpretation of radiographs, it remains essential to strengthen human reading capacities. The TXS can assist expert readers as reporting standard and less-experienced readers as teaching tool.

5.5 Proposals for future studies

Beside the constant development of the CAD4TB software by its designers, further research to improve its accuracy via integration of patient characteristics is of great interest. A prediction model proposed by Zaidi et al. showed promising results in a small scale study, but has not yet been reported in detail and lacks external validation (Zaidi et al. 2014). The development and testing of different approaches to combine clinical and radiological characteristics into one probability score may eventually yield a more accurate screening and triage test that can fully meet the WHO’s high-priority target product profile requirements (WHO 2014c).

The accuracy of CAD4TB to detect pulmonary tuberculosis has been validated in different sub-Saharan settings. Prospective evaluation studies of targeted active case finding strategies with CAD4TB as initial test should follow. Particular interest lies in the influence of CAD4TB on case detection rates, operational aspects and cost-effectiveness.

An evaluation of CAD4TB on public datasets should be pursued. Ideally, these datasets stem from well-characterised patients and include at least microbiological results, better yet, clinical and radiological data, too. Already existing datasets that have been used for derivation and validation of automated reading solutions should be published.

Future research should focus on the compilation of a reference image set and universal consensus on reading terminology.

The TXS needs to be simplified to benefit non-expert readers. Statistically remodelling of the TXS by uni- and multivariate analysis may reveal non-predictive features that can be omitted. After training for the recognition of the remaining features, the performance of non-experts using the remodelled TXS should be reassessed.

6 Conclusion

We evaluated two efforts, the CAD4TB and TXS, to enhance the reproducibility and accuracy of chest radiograph interpretation for the presence of pulmonary tuberculosis.

The computer-aided diagnosis system CAD4TB proved as an accurate and reproducible test for the detection of culture-positive PTB on radiographs of symptomatic patients without the need for trained reading personnel. It detected PTB significantly better in smear-positive over smear-negative patients and in HIV-negative compared to HIV-positive patients. CAD4TB was as accurate as two expert readers and one clinical officer for the question of ‘any abnormality’ on the chest radiograph. The software significantly outperformed the clinical officer, but did not reach the accuracy of both expert readers for tuberculosis specific reading thresholds. Its very condensed output suggests its use as a triage or screening test.

The TXS, a standardized reporting and scoring form to assist human readers, yielded less conclusive results. Its descriptive part together with a simple four-point categorical conclusion code proved valuable as reporting standard for expert readers. Via an accurate tabulation of features seen, assignment of scores and their translation into conclusion categories, the TXS unveiled the decision making process of expert readers. Together with a reference set of images, the TXS may prove effective in the training of non-expert readers. However, the scoring function did enhance neither the diagnostic accuracy nor the reproducibility of expert readings and was detrimental to our clinical officer’s diagnostic performance. This finding prompts additional research on its improvement and let us discourage less experienced readers from its use.

In conclusion, our study provides evidence on the performance of different CXR interpretation modalities to facilitate the effective utilisation of chest radiography as a triage, screening and diagnostic test for PTB. While “fully automating the chest exam” (Conners et al. 1982) might be possible eventually, in its current state, CAD4TB can provide an efficient alternative to select individuals at highest risk for PTB in screening and triage situations. However, it cannot interpret a chest radiograph on par with an experienced human reader. Until automated reading of CXRs advances considerably and in view of the anticipated scale-up of digital radiography in resource-constrained high-burden settings thanks to innovative technology, skilled readers are needed more than ever. The TXS as reporting standard and teaching tool may facilitate quality assurance of CXR interpretation and the training of human readers.

Effective treatment for most patients with tuberculosis is available, nevertheless 1.5 million people died from the disease in 2014 (WHO 2015a). A major bottleneck in TB control is the lack of access to accurate and rapid diagnosis. In the absence of a rapid point-of-care test that can detect all forms of tuberculosis, we are well advised to make the most of existing diagnostic methods to reach, treat and cure an estimated 3.6 million people living with undiagnosed TB.

Danksagung

Herzlichen Dank möchte ich Dr. Klaus Reither für die Ideengebung und ausgezeichnete Betreuung aussprechen. Ohne seine Unterstützung und Zuversicht wäre diese Arbeit nicht entstanden.

Ebenso möchte ich Professor Dirk Wagner herzlich dafür danken, dass er bereit war dieses Forschungsvorhaben von Beginn an bis zu guter Letzt zu begleiten, zu fördern und zu bereichern.

Besonderer Dank gilt Amanda Ross, durch deren statistische Hilfestellung, unvoreingenommenen Blick und konstruktive Kritik die Arbeit an Güte gewann.

Besten Dank möchte ich Professor Bram van Ginneken und Rick Philipsen für die fruchtbare und stets zuverlässige Zusammenarbeit aussprechen.

Mein herzlicher Dank gilt Dr. Claudia Daubenberger für das in mich gesetzte Vertrauen, das Kontakteknüpfen und die klaren Worte im richtigen Moment.

Dr. Jerry Hella, Dr. Francis Mhimbira, Dr. Andreas Steiner und Dr. Levan Jugheli danke ich für ihre Hilfe bei Organisation und Extraktion der Daten. Bei Mwinyikambi Salum, Dr. Jan van den Hombergh und Dr. Jaffer Dharsee möchte ich mich für die Auswertung der Röntgenbilder bedanken. Ich möchte mich zudem herzlich bei allen Mitarbeitern der Tuberkulose-Klinik des Ifakara Health Instituts in Bagamoyo bedanken. Sie alle hatten daran teil meinen Forschungsaufenthalt zu einer beruflich und persönlich kostbaren Zeit zu gestalten.

Mein besonderer Dank gilt zudem allen Patienten, die bereit waren an der TB Cohort und TB CHILD Studie teilzunehmen.

Die TB Cohort Studie wurde durch die Rudolf Geigy Stiftung, Schweiz, gefördert. Die TB CHILD Studie wurde durch die European & Developing Countries Clinical Trials Partnership (EDCTP) als Teil des Projekts ‘Evaluation of new and emerging diagnostics for childhood tuberculosis in high burden countries’ (IP.2009.32040.007) ermöglicht.

Asante sana möchte ich Rama N’kane und Fundi Raimon sagen. Ohne Tisch lässt sich nun mal keine Dissertation schreiben.

Zu guter Letzt möchte ich meiner Familie, insbesondere meinen Eltern Lilli und Walter Breuninger, frei nach J. W. von Goethe danken: für Flügel und Wurzeln.

Appendix

A Inclusion criteria for the CAD4TB and TXS validation study, the TB Cohort and TB CHILD study

CAD4TB and TXS validation study:

Inclusion criteria:

- patients with persistent cough for ≥ 2 weeks and at least one of the following TB associated findings:

  haemoptysis, chest pain, fever, night sweats, constant fatigue, recent unexplained weight loss, loss of appetite, malaise, contact with TB case
- > 18yrs
- signed informed consent / witnessed oral consent to participate in TB CHILD /TB Cohort Study

TB Cohort study:

Inclusion criteria:

- patients who have clinical signs and symptoms suggestive of pulmonary TB
- for pulmonary TB: persistent cough for ≥ 2 weeks and at least one of the following TB associated findings: haemoptysis, chest pain, fever, night sweats, constant fatigue, recent unexplained weight loss, loss of appetite
- for extrapulmonary TB: suspected tuberculosis of organs other than the lungs, such as lymph nodes, abdomen, genitourinary tract, skin, joints and bones, meninges, or others
- patients older than 4 weeks of age
- any patient attending the NTLP clinic who gives informed consent to participate in the IHI TB epidemiology study
- patients residing within the study areas and who are not planning to move from the study area within the 18 month period following their inclusion into the study

Exclusion criteria:

- patients who do not agree to participate in the study or whose legal guardian does not agree that they participate
- patients who are non-residents of the study areas
- severely sick patients

TB CHILD study:

Inclusion criteria:

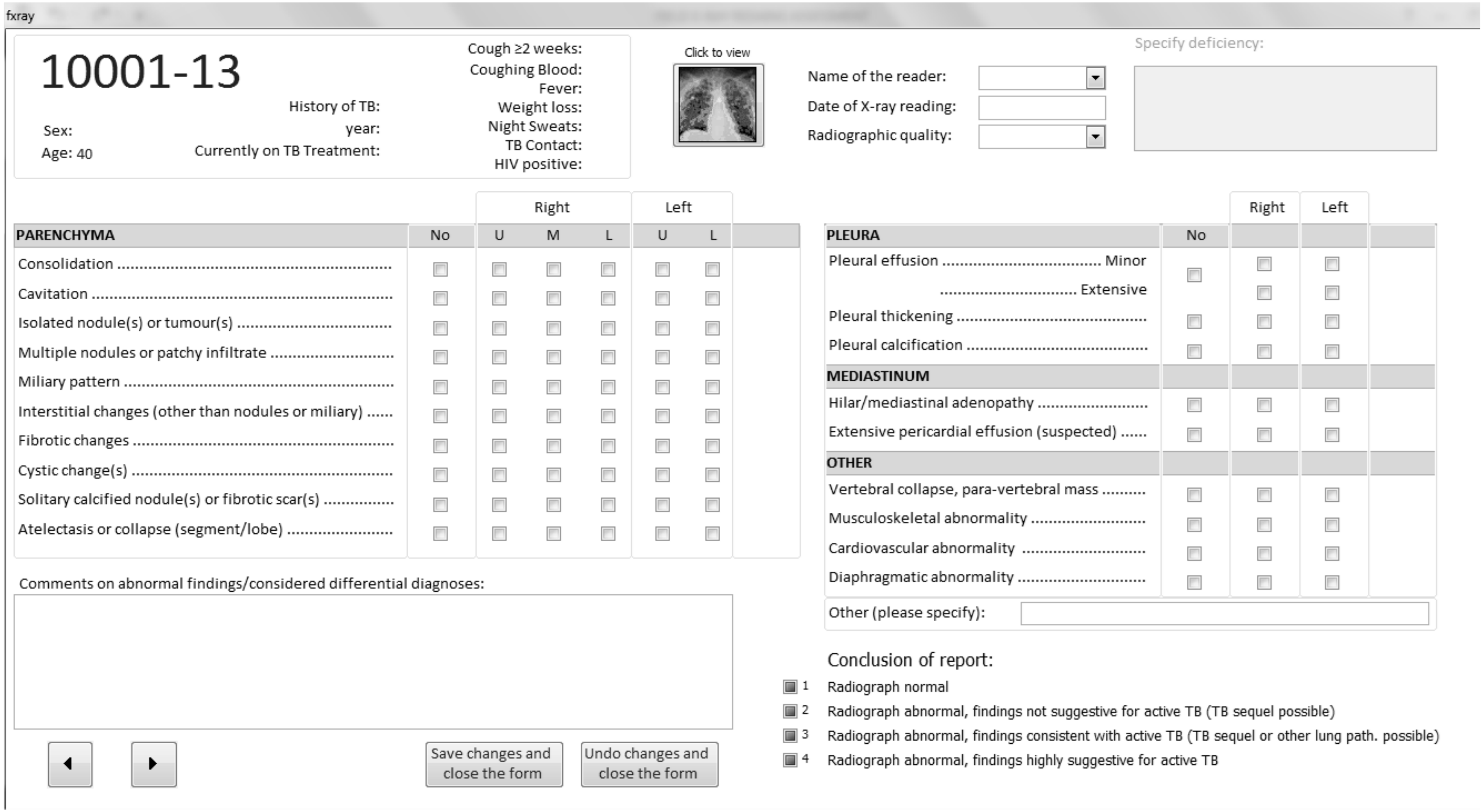
- signed informed consent form / witnessed oral consent
- >18yrs

- persistent cough for ≥ 2 weeks <u>and</u> at least one of the following conditions: haemoptysis, chest pain, fever, night sweats, malaise, unexplained weight loss within the last 3 months, loss of appetite, contact with TB case

Exclusion criterion:

- TB treatment in the past year

B Tanzanian Chest X-ray Score



fxray

10001-13

Sex:

Age: 40

History of TB:

year:

Currently on TB Treatment:

Cough ≥2 weeks:

Coughing Blood:

Fever:

Weight loss:

Night Sweats:

TB Contact:

HIV positive:

Name of the reader:

Date of X-ray reading:

Radiographic quality:

Specify deficiency:

		Right			Left	
PARENCHYMA	No	U	M	L	U	L
Consolidation	□	□	□	□	□	□
Cavitation	□	□	□	□	□	□
Isolated nodule(s) or tumour(s)	□	□	□	□	□	□
Multiple nodules or patchy infiltrate	□	□	□	□	□	□
Miliary pattern	□	□	□	□	□	□
Interstitial changes (other than nodules or miliary)	□	□	□	□	□	□
Fibrotic changes	□	□	□	□	□	□
Cystic change(s)	□	□	□	□	□	□
Solitary calcified nodule(s) or fibrotic scar(s)	□	□	□	□	□	□
Atelectasis or collapse (segment/lobe)	□	□	□	□	□	□

		Right	Left
PLEURA	No		
Pleural effusion .... Minor	□	□	□
.... Extensive		□	□
Pleural thickening	□	□	□
Pleural calcification	□	□	□
MEDIASTINUM			
Hilar/mediastinal adenopathy	□	□	□
Extensive pericardial effusion (suspected)	□	□	□
OTHER			
Vertebral collapse, para-vertebral mass	□	□	□
Musculoskeletal abnormality	□	□	□
Cardiovascular abnormality	□	□	□
Diaphragmatic abnormality	□	□	□

Other (please specify):

Comments on abnormal findings/considered differential diagnoses:

Conclusion of report:

■ 1 Radiograph normal

■ 2 Radiograph abnormal, findings not suggestive for active TB (TB sequel possible)

■ 3 Radiograph abnormal, findings consistent with active TB (TB sequel or other lung path. possible)

■ 4 Radiograph abnormal, findings highly suggestive for active TB

◀
▶
Save changes and close the form
Undo changes and close the form

Figure 16 Screenshot of the user interface of the digitalised Tanzanian Chest X-ray Score (TXS)

C Agreement tables

expert 1		TXS				total
		1	2	3	4	
manual	1	396	21	2	3	422
	2	3	68	26	22	119
	3	1	11	62	85	159
	4	0	2	21	138	161
total		400	102	111	248	861

expert 2		TXS				total
		1	2	3	4	
manual	1	325	33	0	0	358
	2	0	78	26	13	117
	3	1	32	76	98	207
	4	4	4	10	161	179
total		330	147	112	272	861

clinical officer		TXS				total
		1	2	3	4	
manual	1	77	6	3	0	86
	2	3	45	220	70	338
	3	2	29	58	311	400
	4	2	0	0	35	37
total		84	80	281	416	861

Table 13 Agreement between manual and TXS conclusion

manual		expert2				total
		1	2	3	4	
expert1	1	326	42	49	5	422
	2	24	39	51	5	119
	3	8	26	81	44	159
	4	0	10	26	125	161
total		358	117	207	179	861

manual		clinical officer				total
		1	2	3	4	
expert1	1	82	223	108	9	422
	2	2	41	71	5	119
	3	2	51	95	11	159
	4	0	23	126	12	161
total		86	338	400	37	861

manual		clinical officer				total
		1	2	3	4	
expert2	1	76	193	84	5	358
	2	5	37	69	6	117
	3	5	83	110	9	207
	4	0	25	137	17	179
total		86	338	400	37	861

Table 14 Inter-reader agreement for the manual conclusion

TXS		expert2				total
		1	2	3	4	
expert1	1	294	68	29	9	400
	2	20	42	21	19	102
	3	6	13	27	65	111
	4	10	24	35	179	248
total		330	147	112	272	861

TXS		clinical officer				total
		1	2	3	4	
expert1	1	77	43	154	126	400
	2	4	13	30	55	102
	3	0	5	40	66	111
	4	3	19	57	169	248
total		84	80	281	416	861

TXS		clinical officer				total
		1	2	3	4	
expert2	1	73	33	128	96	330
	2	7	16	46	78	147
	3	3	10	39	60	112
	4	1	21	68	182	272
total		84	80	281	416	861

Table 15 Inter-reader agreement for the TXS conclusion

manual		expert1				total
		1	2	3	4	
expert1	1	72	8	4	1	85
	2	8	14	9	0	31
	3	6	8	21	8	43
	4	1	0	15	24	40
total		87	30	49	33	199

manual		clinical officer				total
		1	2	3	4	
clinical officer	1	3	10	0	0	13
	2	2	55	24	0	81
	3	2	30	60	1	93
	4	0	2	10	0	12
total		7	97	94	1	199

TXS		expert1				total
		1	2	3	4	
expert1	1	66	12	3	0	81
	2	6	12	3	2	23
	3	5	4	14	8	31
	4	4	5	13	42	64
total		81	33	33	52	199

TXS		clinical officer				total
		1	2	3	4	
clinical officer	1	3	3	2	6	14
	2	0	2	4	11	17
	3	3	3	21	39	66
	4	2	4	18	78	102
total		8	12	45	134	199

Table 16 Intra-reader agreement for the manual and TXS conclusion

D Overview CAD4TB studies

author **year** **software version**	**setting, study population**	**human readers & reading methodology**	**refer-ence**	**CAD4TB accuracy**	**performance human readers vs. CAD4TB**	**additional findings**
Steiner 2015 **v3.07**	n=511, asymptomatic prisoners, Tanzania *1st real world screening setting evaluation*	7 health care professionals [4 cat.]	radio.	**Az=0.75**	2 * worse, 2 * better than, 3 comparable to CAD4TB	CAD4TB without diagnostic delay (>99%) vs. 12.2% delay >24h with conventional reading
Philipsen 2015 **v3.07**	n=388, presumptive TB patients (33% HIV+), South Africa *ACR as triage for Xpert*	2 CRRS-certified “B”-reader, 1 specialist reader [0-100]	bact.	**Az=0.79**	Az=0.76-0.81 specialist readers comparable to CAD4TB	a CAD4TB score ≥ 85 as triage test increases Xpert throughput from 45 to 113 per day, reduces costs per screened subject from $13.09 to $6.72 and costs per notified TB case from $90.70 to $54.34 with a compromise in sensitivity from 78.9% (Xpert alone) to 67.6%
Breuninger 2014 **v3.07**	n=566, presumptive TB patients (43% HIV+), Tanzania. *clinical validation*	1 expert reader & 1 CO [4 cat.]	bact.	**Az=0.81**	comparable for ‘any abnormality’, for TB specific thresholds: CO * worse, expert * better	CAD4TB * less accurate in smear- and HIV+ patients
Muyoyeta 2014 **v1.08**	n=350, presumptive TB patients (54% HIV+), Zambia *1st prospective clinical setting evaluation*	-	Xpert	**Az=0.71**	-	

Khan 2014 **v3.07**	n=186, presumptive TB patients, Pakistan *active case finding* *at health care facilities*	2 radiologists and 1 CO	radio.	sens.78% spec.79%	-	a CAD4TB score ≥ 80 detected high proportion of non-TB abnormalities according to review by humans
Zaidi 2014 **v3.07**	n=324, presumptive TB patients, Pakistan *CAD4TB + variables*	-	Xpert	**-**	-	prediction model incl. CAD4TB, cough>2weeks, age & gender: **Az=0.87**
Maduskar 2013 **v1.08**	n=161, (specimen bank of) presumptive TB patients (68% HIV+), Zambia *validation*	4 CO [0-100]	bact. / radio.	**Az= 0.73 / 0.91**	Az=0.65-0.75 / 0.89-0.94 comparable (except 1 CO * worse for bact.reference)	
Ginneken 2012 **v1.08**	n=100, presumptive TB patients, sub-Saharan Africa.	7 inexperienced & 1 experienced observer [0-100]	bact.	**Az=0.82**	expert Az=0.84, non-experts: Az=0.69-0.86	independent combination (averaging) of human and CAD4TB score increased performance of all readers: expert=0.85, non-experts=0.73-0.87 (in 4/7 non-experts *)
Hogeweg 2011 **CADx**	n=95, CXRs from screening of high risk group, UK *screening* *(preselected image set)*		bact.	**Az=0.86**		at sens 95% -> spec 60%

CO = clinical officer, * = significantly

Table 17 Overview of CAD4TB validation studies

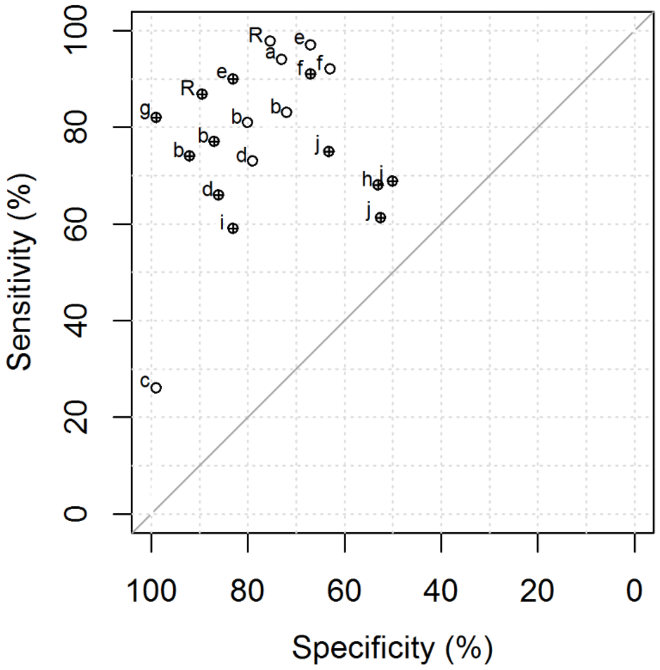
E Overview reproducibility and diagnostic accuracy of different CXR reading methodologies

author year	**setting**	**sam-ple size n**	**refer-ence stand-ard**	**κ**		**sens [%]**	**spec [%]**	**HIV cases only**	**type of readers**	**abnor-mality type**	**reading method**
				inter	intra						
Waitt 2013	tertiary referral hospital, Malawi	60	bact.	-		61 69 75	53 50 63	no	2 clinical officers before & after CRRS-course	TB	- /TIRS /CRRS
Pinto 2013	university hospital, ZA	473	bact	0.52 /-		93 86	28 64	no	2 experts (specialist physicians)	TB / ≥2	CRRS /score
Story 2012	high-risk group screening, UK	47510	bact.	-		82	99	no	radiographers	TB	5 cat.
van't Hoog 2012	prevalence survey, Kenya	20566	bact.	-		94	73	no	clinical officers	any	-
van't Hoog 2011	prevalence survey, Kenya	1143	bact.	(n=1031 neg. CXRs) 0.52/0.40 (any/TB)* (n=112 PTB CXRs) 0.82/0.74 (any/TB)* (n=655) #any 0.57-0.63	(n≈200) #any 0.43-0.60	77, 74* 83, 81* 95#	87, 92* 72, 80* 73#	no	2 experts***** (radio+pulm), 3 clinical officers#	TB any	CRRS***** - #
Dawson 2010	ART service, ZA	203	bact.	0.63		68	53	yes	2 experts (resp+infec) (CRRS-certified)	TB	CRRS
Lewis 2009	Miners screening, ZA	1955	bact.	-		26	99	no	experts (radio)	any	-
Shah 2009	HIV clinic, Ethiopia	438	bact.	0.61		59	83	yes	experts (radio)	TB	standard form, 5 cat.

Day 2006	HIV clinic for miners, ZA	899	bact. &clin/ rad.	-		66 73	86 79	yes	doctor	TB any	-
den Boon, 2006	prevalence survey, ZA	1170	bact. (s/c)	-		90 97	83 67	no	expert (pulm)	TB any	CRRS
Zellweger 2006	Immigrant screening, Switzerland	377	radio.	all 3 readers: 0.56 experts only: 0.85	0.76, 0.90, 0.39	-	-	no	2 experts (chest), 1 junior doctor	TB	4 cat.
den Boon 2005	prevalence survey, ZA	810	-	0.69 0.47	0.90 0.85 (n=104)	-	-	no	expert (pulm) , grade ‘A’ reader (UICC/ILO)	TB any	CRRS
Balabanova 2005	general clinic, Russia	50	radio.	0.39 0.38° 0.45^ 0.39’	0.49 0.48° 0.53^ 0.48’ (n=10)	-	-	no	101 physicians: -61 TB specialists° -25 radiologists^ -15 resp. specialists’	TB	structured questionnaire
van Cleef 2005	Chest clinic, Kenya	998	bact.	0.55 (overall), 0.75 (TB) (n=714)		91 92	67 63	no	radiologists	TB any	4 cat.
Graham 2002	immigrant screening, Canada	973	radio.	0.45 0.56	0.59 0.72 (10%)	-	-	no	radiologists	TB any	5 cat.

Table 18 Overview CXR performance for TB detection: reproducibility and diagnostic accuracy, table modified after Maduskar et al. 2013b

F Comparison of study results and literature findings



Legend.

a	(Hoog et al. 2012)
b	(Hoog et al. 2011b)
c	(Lewis et al. 2009)
d	(Day et al. 2006)
e	(Den Boon et al. 2006)
f	(van Cleeff et al. 2005)
g	(Story et al. 2012)
h	(Dawson et al. 2010)
i	(Shah et al. 2009)
j	(Waitt et al. 2013)
R	Review (Hoog et al. 2014)

Figure 17 Diagnostic accuracy values of CXR for PTB detection as reported in the literature

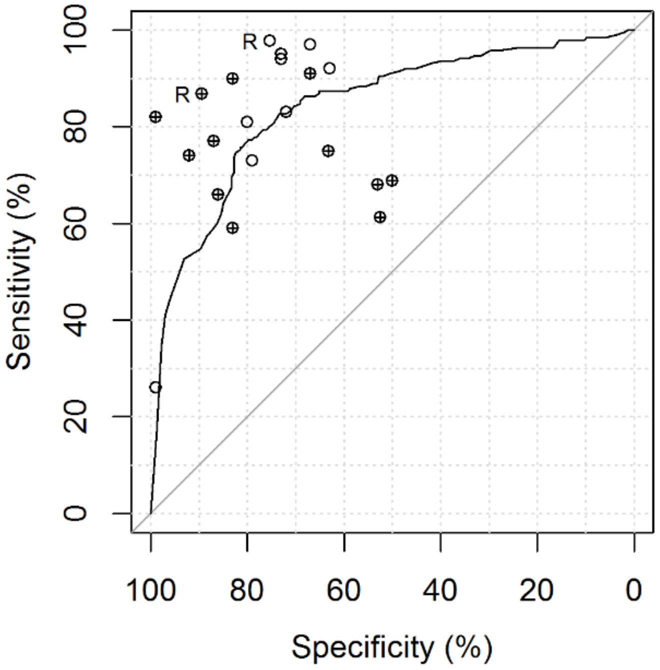


Figure 18 CAD4TB performance in this study vs. literature findings

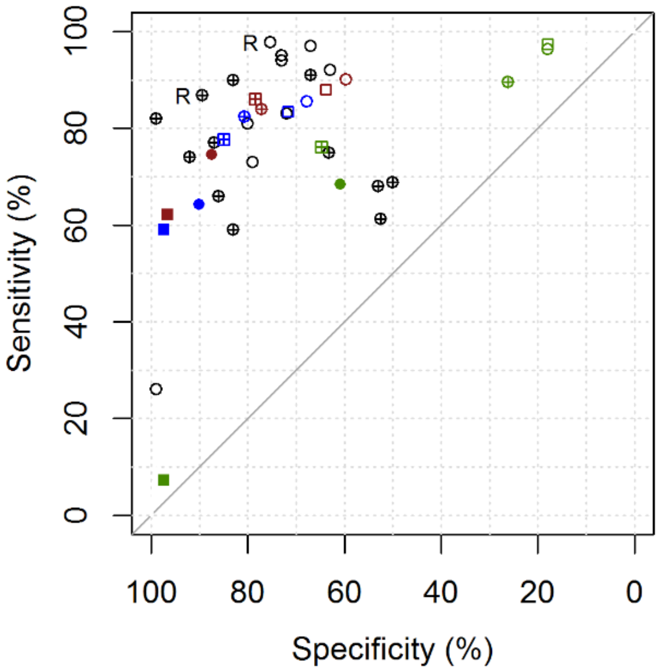


Figure 19 Human reading results (TXS conclusion & manual conclusion) in this study vs. literature findings

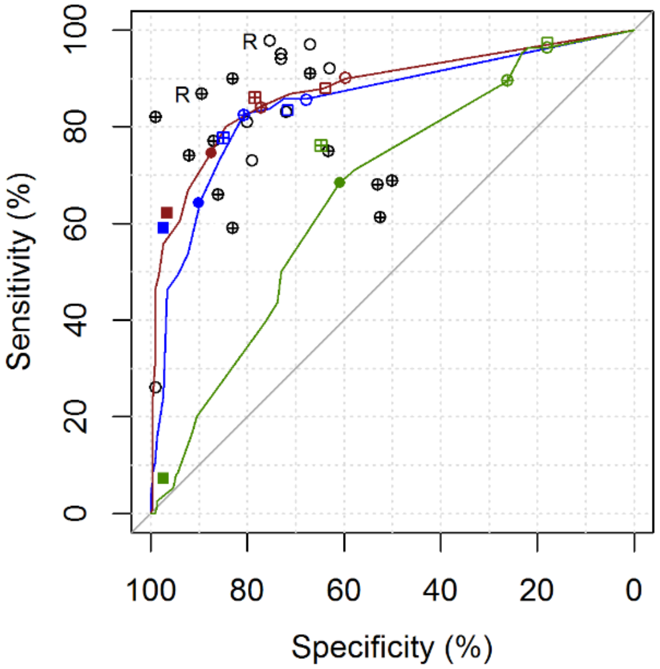


Figure 20 Human reading results (TXS conclusion + continuous output & manual conclusion) in this study vs. literature findings

Legend.

CAD4TB	—					
TXS continuous (= ROC curve)	—	expert reader 1	—	expert reader 2	—	clinical officer
TXS conclusion	○○○	any abnormality	⊕⊕⊕	TB consistent abnormalities	●●●	abnormalities highly suggestive for TB
manual conclusion	□□□		⊞⊞⊞		■■■	

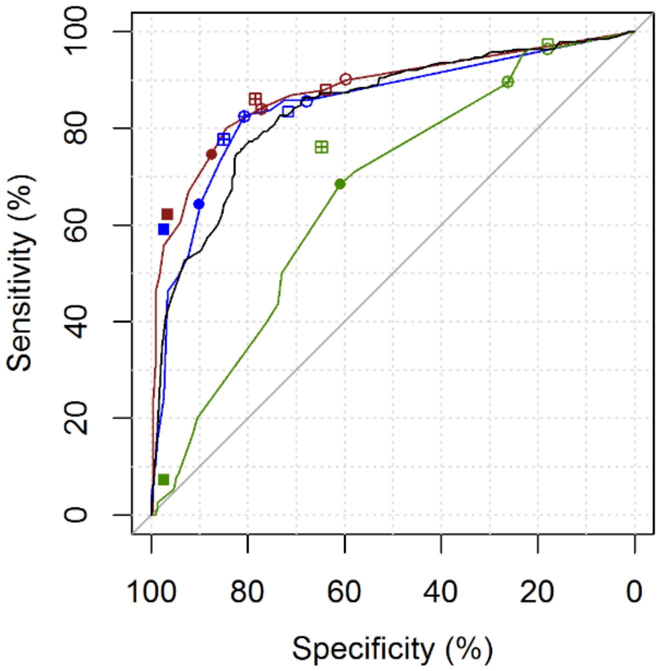


Figure 21 Diagnostic accuracy of CAD4TB and human reading in this study

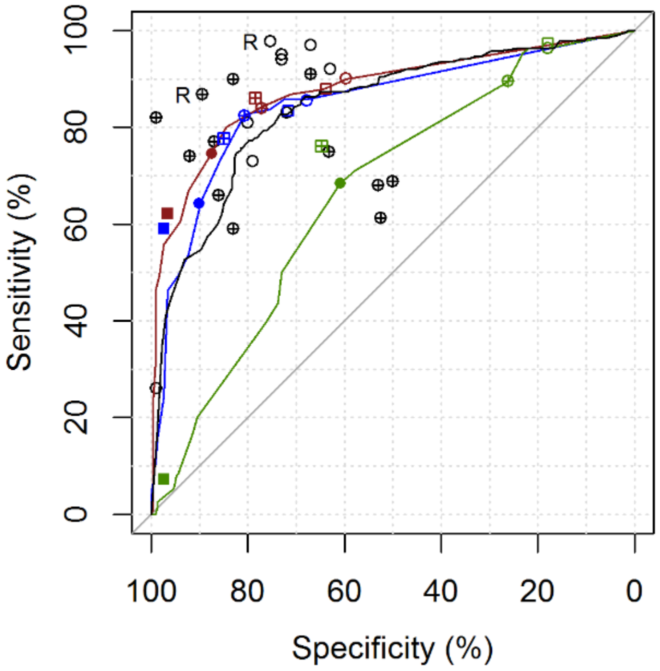
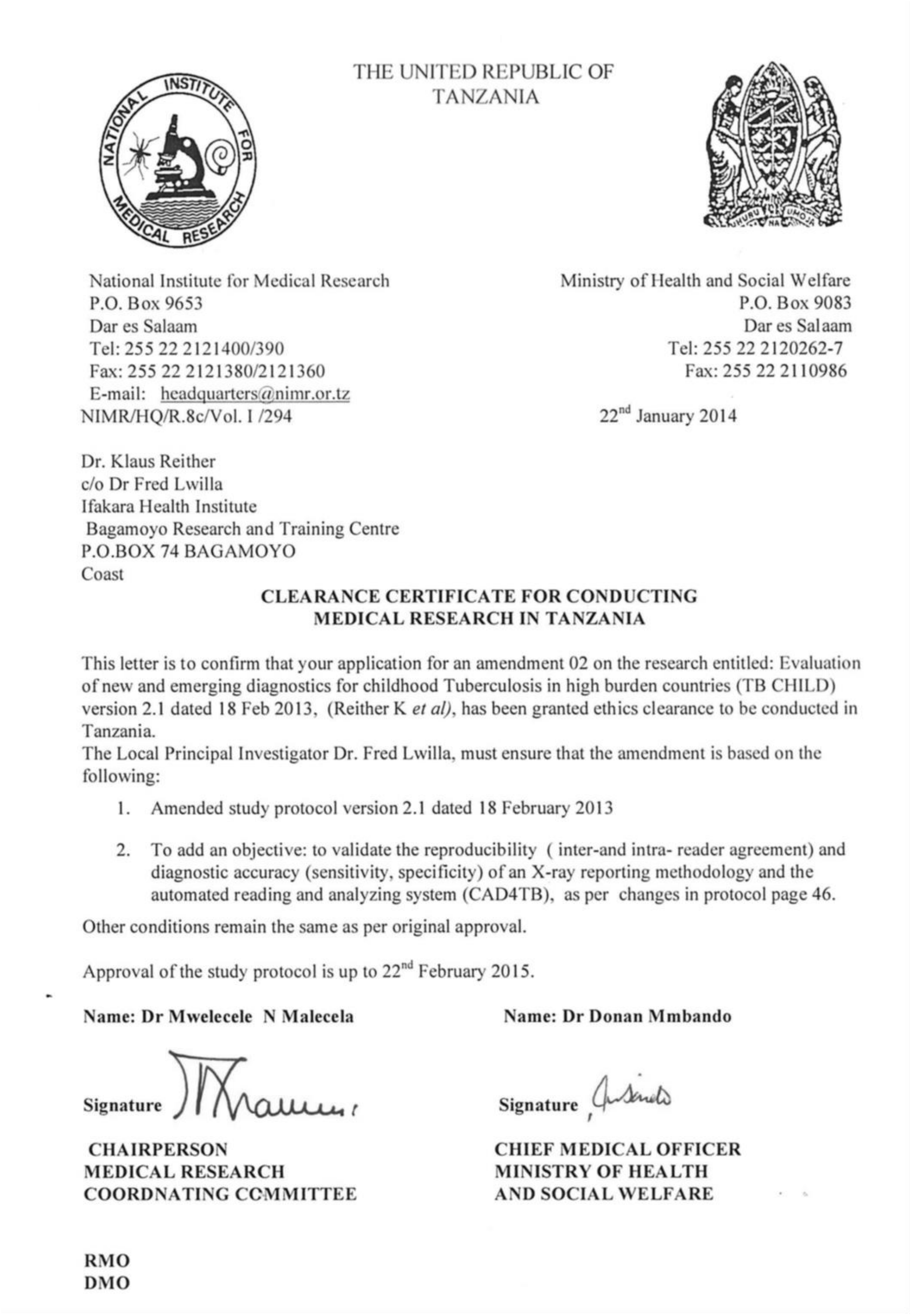


Figure 22 CAD4TB and human reading (TXS conclusion + continuous output & manual conclusion) in this study vs. literature findings

G Ethical approval



THE UNITED REPUBLIC OF
TANZANIA

National Institute for Medical Research
P.O. Box 9653
Dar es Salaam
Tel: 255 22 2121400/390
Fax: 255 22 2121380/2121360
E-mail: headquarters@nimr.or.tz

Ministry of Health and Social Welfare
P.O. Box 9083
Dar es Salaam
Tel: 255 22 2120262-7
Fax: 255 22 2110986

NIMR/HQ/R.8c/Vol. I /294

22nd January 2014

Dr. Klaus Reither
c/o Dr Fred Lwilla
Ifakara Health Institute
Bagamoyo Research and Training Centre
P.O.BOX 74 BAGAMOYO
Coast

CLEARANCE CERTIFICATE FOR CONDUCTING MEDICAL RESEARCH IN TANZANIA

This letter is to confirm that your application for an amendment 02 on the research entitled: Evaluation of new and emerging diagnostics for childhood Tuberculosis in high burden countries (TB CHILD) version 2.1 dated 18 Feb 2013, (Reither K *et al)*, has been granted ethics clearance to be conducted in Tanzania.

The Local Principal Investigator Dr. Fred Lwilla, must ensure that the amendment is based on the following:

1. Amended study protocol version 2.1 dated 18 February 2013
2. To add an objective: to validate the reproducibility (inter-and intra- reader agreement) and diagnostic accuracy (sensitivity, specificity) of an X-ray reporting methodology and the automated reading and analyzing system (CAD4TB), as per changes in protocol page 46.

Other conditions remain the same as per original approval.

Approval of the study protocol is up to 22nd February 2015.

Name: Dr Mwelecele N Malecela

Signature

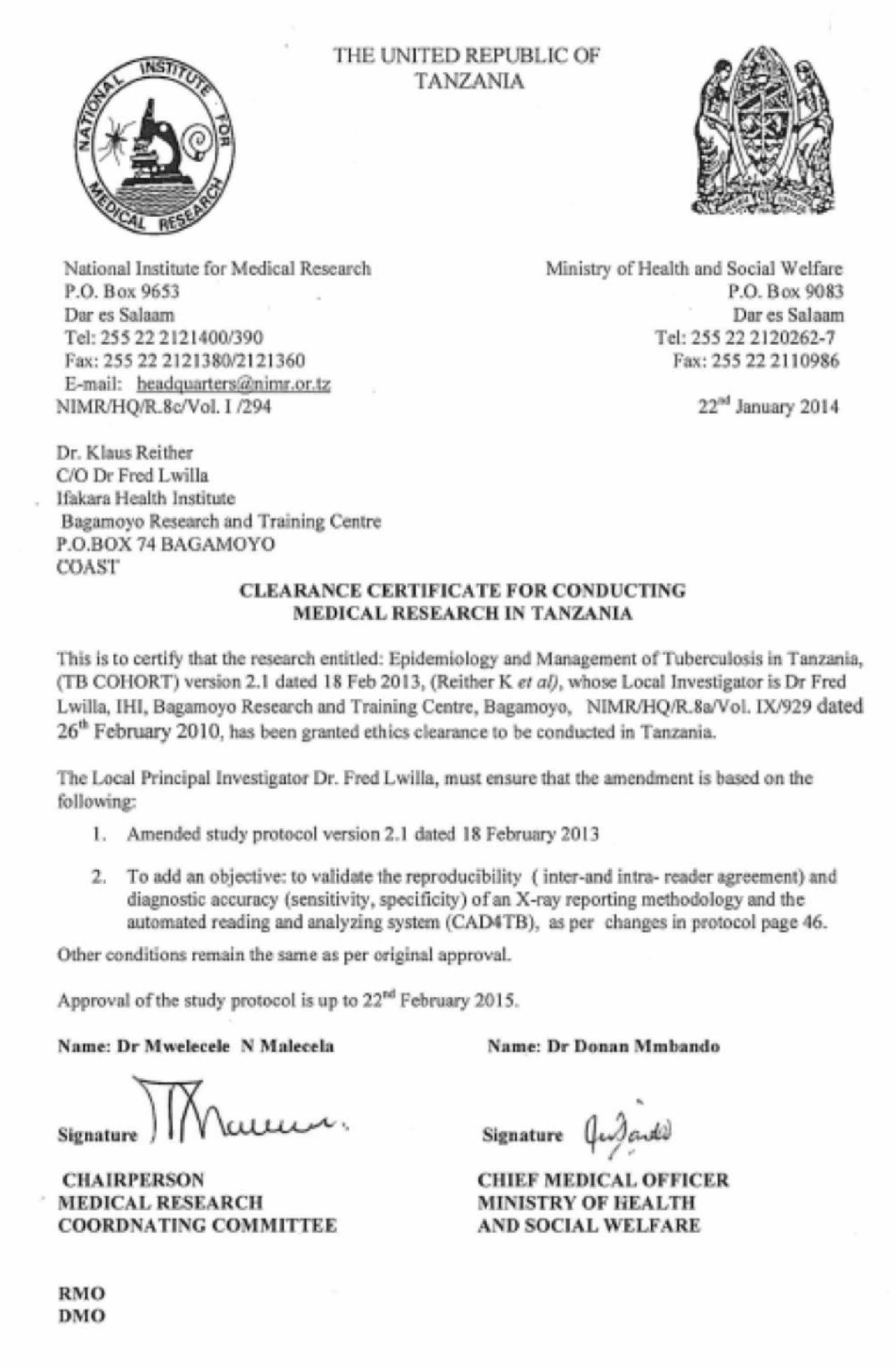
CHAIRPERSON
MEDICAL RESEARCH
COORDNATING COMMITTEE

Name: Dr Donan Mmbando

Signature

CHIEF MEDICAL OFFICER
MINISTRY OF HEALTH
AND SOCIAL WELFARE

RMO
DMO



THE UNITED REPUBLIC OF
TANZANIA

National Institute for Medical Research
P.O. Box 9653
Dar es Salaam
Tel: 255 22 2121400/390
Fax: 255 22 2121380/2121360
E-mail: headquarters@nimr.or.tz
NIMR/HQ/R.8c/Vol. I /294

Ministry of Health and Social Welfare
P.O. Box 9083
Dar es Salaam
Tel: 255 22 2120262-7
Fax: 255 22 2110986

22nd January 2014

Dr. Klaus Reither
C/O Dr Fred Lwilla
Ifakara Health Institute
Bagamoyo Research and Training Centre
P.O.BOX 74 BAGAMOYO
COAST

CLEARANCE CERTIFICATE FOR CONDUCTING MEDICAL RESEARCH IN TANZANIA

This is to certify that the research entitled: Epidemiology and Management of Tuberculosis in Tanzania, (TB COHORT) version 2.1 dated 18 Feb 2013, (Reither K *et al)*, whose Local Investigator is Dr Fred Lwilla, IHI, Bagamoyo Research and Training Centre, Bagamoyo, NIMR/HQ/R.8a/Vol. IX/929 dated 26th February 2010, has been granted ethics clearance to be conducted in Tanzania.

The Local Principal Investigator Dr. Fred Lwilla, must ensure that the amendment is based on the following:

1. Amended study protocol version 2.1 dated 18 February 2013
2. To add an objective: to validate the reproducibility (inter-and intra- reader agreement) and diagnostic accuracy (sensitivity, specificity) of an X-ray reporting methodology and the automated reading and analyzing system (CAD4TB), as per changes in protocol page 46.

Other conditions remain the same as per original approval.

Approval of the study protocol is up to 22nd February 2015.

Name: Dr Mwelecele N Malecela

Signature

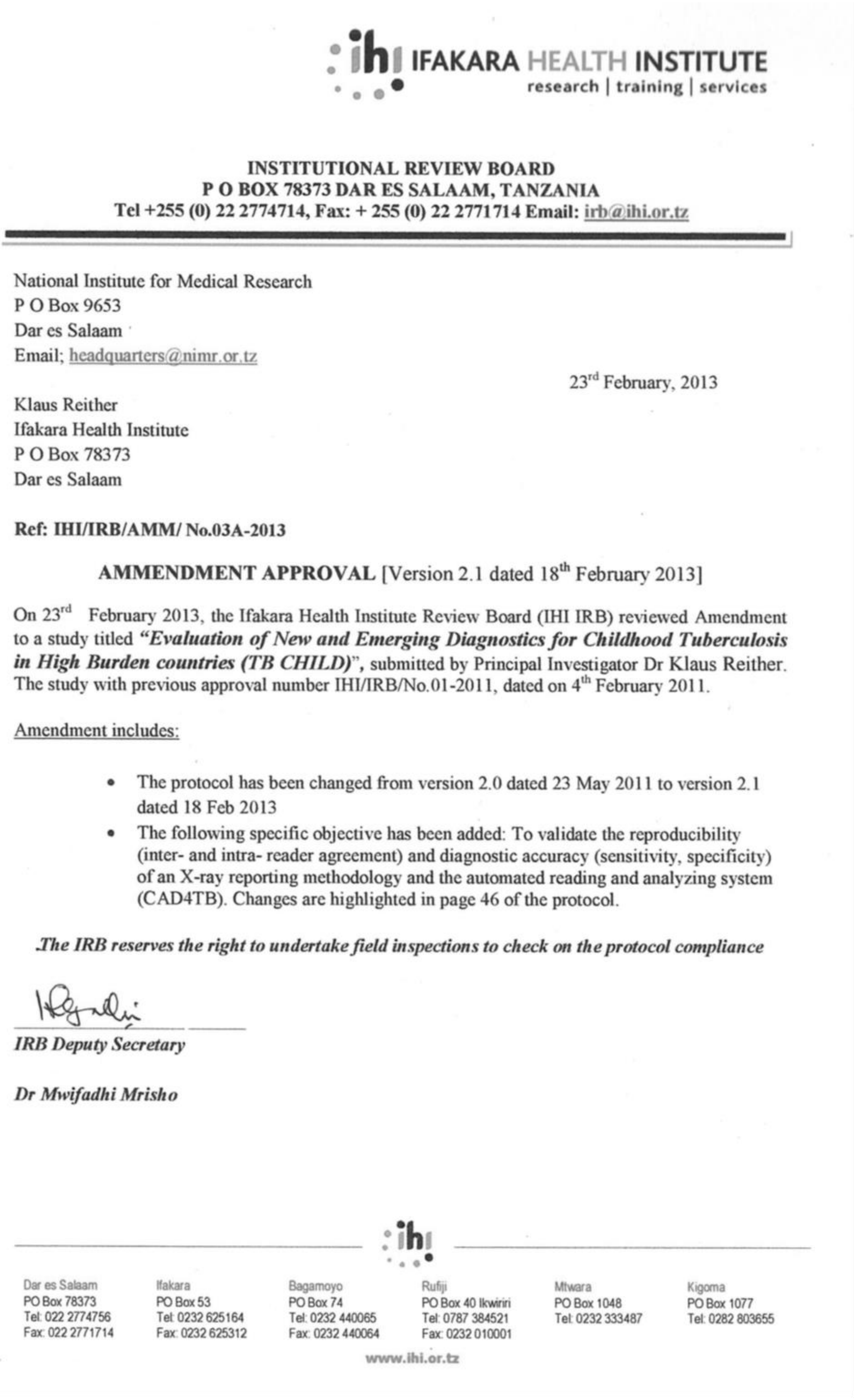
CHAIRPERSON
MEDICAL RESEARCH
COORDNATING COMMITTEE

Name: Dr Donan Mmbando

Signature

CHIEF MEDICAL OFFICER
MINISTRY OF HEALTH
AND SOCIAL WELFARE

RMO
DMO



IFAKARA HEALTH INSTITUTE
research | training | services

INSTITUTIONAL REVIEW BOARD
P O BOX 78373 DAR ES SALAAM, TANZANIA
Tel +255 (0) 22 2774714, Fax: + 255 (0) 22 2771714 Email: irb@ihi.or.tz

National Institute for Medical Research
P O Box 9653
Dar es Salaam
Email; headquarters@nimr.or.tz

23rd February, 2013

Klaus Reither
Ifakara Health Institute
P O Box 78373
Dar es Salaam

Ref: IHI/IRB/AMM/ No.03A-2013

AMMENDMENT APPROVAL [Version 2.1 dated 18th February 2013]

On 23rd February 2013, the Ifakara Health Institute Review Board (IHI IRB) reviewed Amendment to a study titled ***"Evaluation of New and Emerging Diagnostics for Childhood Tuberculosis in High Burden countries (TB CHILD)***", submitted by Principal Investigator Dr Klaus Reither. The study with previous approval number IHI/IRB/No.01-2011, dated on 4th February 2011.

Amendment includes:

- The protocol has been changed from version 2.0 dated 23 May 2011 to version 2.1 dated 18 Feb 2013
- The following specific objective has been added: To validate the reproducibility (inter- and intra- reader agreement) and diagnostic accuracy (sensitivity, specificity) of an X-ray reporting methodology and the automated reading and analyzing system (CAD4TB). Changes are highlighted in page 46 of the protocol.

.The IRB reserves the right to undertake field inspections to check on the protocol compliance

IRB Deputy Secretary

Dr Mwifadhi Mrisho

Dar es Salaam
PO Box 78373
Tel: 022 2774756
Fax: 022 2771714

Ifakara
PO Box 53
Tel: 0232 625164
Fax: 0232 625312

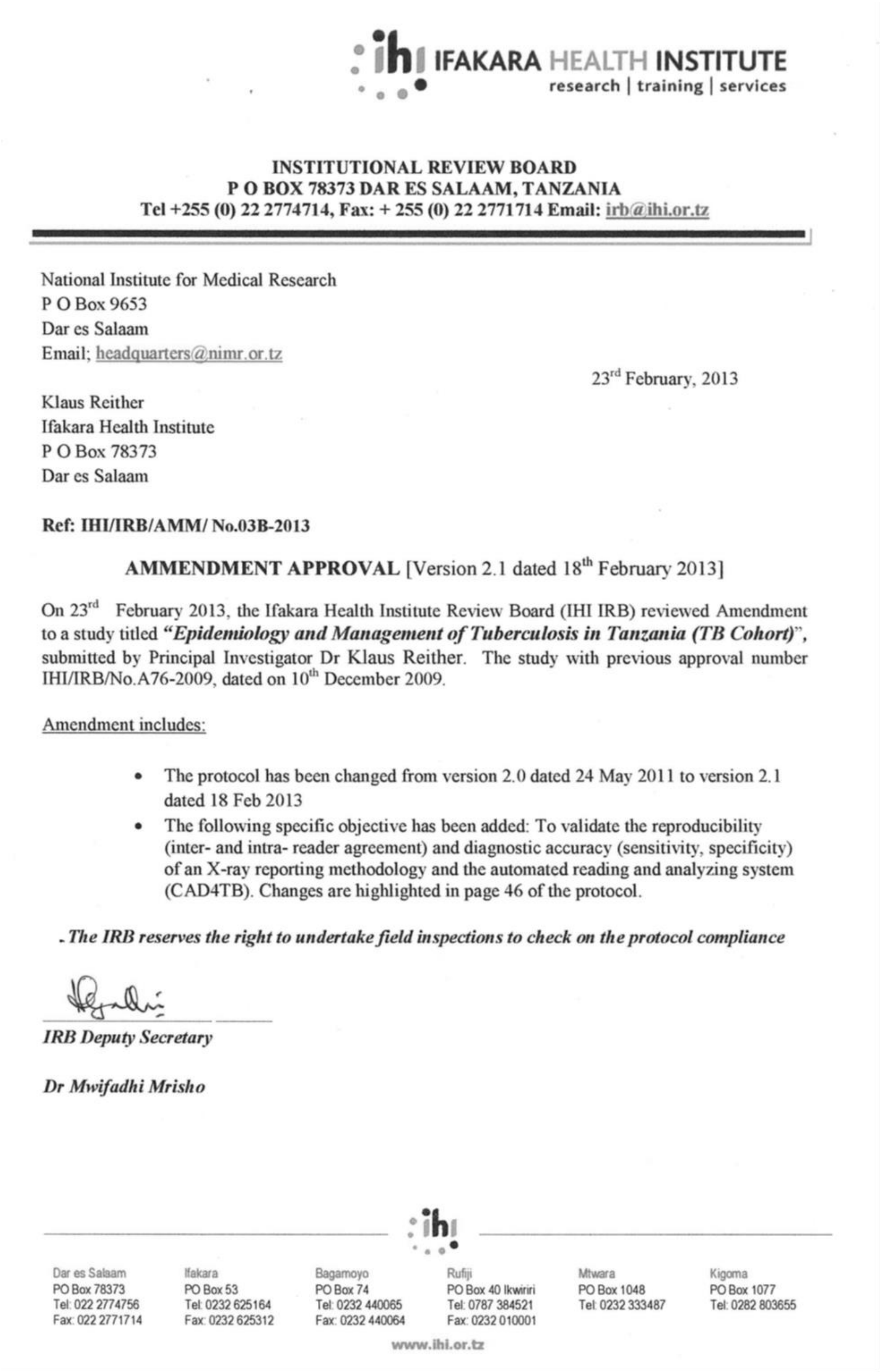
Bagamoyo
PO Box 74
Tel: 0232 440065
Fax: 0232 440064

Rufiji
PO Box 40 Ikwiriri
Tel: 0787 384521
Fax: 0232 010001

Mtwara
PO Box 1048
Tel: 0232 333487

Kigoma
PO Box 1077
Tel: 0282 803655

www.ihi.or.tz



IFAKARA HEALTH INSTITUTE
research | training | services

INSTITUTIONAL REVIEW BOARD
P O BOX 78373 DAR ES SALAAM, TANZANIA
Tel +255 (0) 22 2774714, Fax: + 255 (0) 22 2771714 Email: irb@ihi.or.tz

National Institute for Medical Research
P O Box 9653
Dar es Salaam
Email; headquarters@nimr.or.tz

23rd February, 2013

Klaus Reither
Ifakara Health Institute
P O Box 78373
Dar es Salaam

Ref: IHI/IRB/AMM/ No.03B-2013

AMMENDMENT APPROVAL [Version 2.1 dated 18th February 2013]

On 23rd February 2013, the Ifakara Health Institute Review Board (IHI IRB) reviewed Amendment to a study titled ***"Epidemiology and Management of Tuberculosis in Tanzania (TB Cohort)***", submitted by Principal Investigator Dr Klaus Reither. The study with previous approval number IHI/IRB/No.A76-2009, dated on 10th December 2009.

Amendment includes:

- The protocol has been changed from version 2.0 dated 24 May 2011 to version 2.1 dated 18 Feb 2013
- The following specific objective has been added: To validate the reproducibility (inter- and intra- reader agreement) and diagnostic accuracy (sensitivity, specificity) of an X-ray reporting methodology and the automated reading and analyzing system (CAD4TB). Changes are highlighted in page 46 of the protocol.

- The IRB reserves the right to undertake field inspections to check on the protocol compliance

IRB Deputy Secretary

Dr Mwifadhi Mrisho

Dar es Salaam
PO Box 78373
Tel: 022 2774756
Fax: 022 2771714

Ifakara
PO Box 53
Tel: 0232 625164
Fax: 0232 625312

Bagamoyo
PO Box 74
Tel: 0232 440065
Fax: 0232 440064

Rufiji
PO Box 40 Ikwiriri
Tel: 0787 384521
Fax: 0232 010001

Mtwara
PO Box 1048
Tel: 0232 333487

Kigoma
PO Box 1077
Tel: 0282 803655

www.ihi.or.tz

References

Agizew T, Bachhuber M, Nyirenda S, Makwaruzi V, Tedla Z, Tallaksen R, Parker J, Mboya J, Samandari T (2010) Association of chest radiographic abnormalities with tuberculosis disease in asymptomatic HIV-infected adults. Int J Tuberc Lung Dis 14:324–31

Aliyu G, El-Kamary S, Abimiku A, Brown C, Tracy K, Hungerford L, Blattner W (2013) Prevalence of non-tuberculous mycobacterial infections among tuberculosis suspects in Nigeria. PLoS One 8:e63170

Antani S (2015) Automatic x-ray screening for tuberculosis. SPIE Newsroom. URL http://www.spie.org/x115742.xml (Accessed Jan 03, 2016)

Asmar S, Drancourt M (2015) Rapid culture-based diagnosis of pulmonary tuberculosis in developed and developing countries. Front Microbiol 6:1184

Balabanova Y, Coker R, Fedorin I, Zakharova S, Plavinskij S, Krukov N, Atun R, Drobniewski F (2005) Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. BMJ 331:379–82

Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, Smith PG, Sriram N, Wongsrichanalai C, Linke R, O'Brien R, Perkins M, Cunningham J, Matsoso P, Nathanson CM, Olliaro P, Peeling RW, Ramsay A (2010) Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol 8:S17–S29

Boon S Den, Bateman ED, Enarson D a, Borgdorff MW, Verver S, Lombard CJ, Irusen E, Beyers N, White NW (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 9:1088–96

Boon S Den, White NW, Lill SWP Van, Borgdorff MW, Verver S, Lombard CJ, Bateman ED, Irusen E, Enarson DA, Beyers N (2006) An evaluation of symptom and chest radiographic screening in tuberculosis prevalence surveys. Int J Tuberc Lung Dis 10:876–882

Boyle D, Pai M (2014) UNITAID Tuberculosis Diagnostic Technology and Market Landscape, 3rd Edition.

Boyle D, Pai M (2015) UNITAID Tuberculosis: Diagnostics Technology and Market Landscape, 4th Edition.

Breuninger M, Ginneken B van, Philipsen RHHM, Mhimbira F, Hella JJ, Lwilla F, Hombergh J van den, Ross A, Jugheli L, Wagner D, Reither K (2014) Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa. PLoS One 9:e106381

Breuninger M, Hombergh J van den, Dharsee J, Jugheli L, Hella JJ, Wagner D, Reither K (2014) Tanzanian X-ray score for the detection of active pulmoary TB on chest radiographs: a comparison with subjective assessment. In: Oral Poster Presentation at the 45th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union). Barcelona, Spain, p S556

Buijtels PC a. M, Sande M a. B van der, Graaff CS de, Parkinson S, Verbrugh H a., Petit PLC, Soolingen D van (2009) Nontuberculous Mycobacteria, Zambia. Emerg Infect Dis 15:242–249

Casal M, Gutierrez J, Vaquero M (1997) Comparative evaluation of the mycobacteria growth indicator tube with the BACTEC 460 TB system and Löwenstein-Jensen medium for isolation of mycobacteria from clinical specimens. Int J Tuberc Lung Dis 1:81–84

Chamie G, Luetkemeyer A, Walusimbi-Nanteza M, Okwera A, Whalen CC, Mugerwa RD, Havlir D V, Charlebois ED (2010) Significant variation in presentation of pulmonary tuberculosis across a high resolution of CD4 strata. Int J Tuberc Lung Dis 14:1295–302

Cleeff M van, Kivihya-Ndugga L, Meme H, Odhiambo J, Klatser P (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

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20:37–46

Conners R, Harlow C, Dwyer S (1982) Radiographic image analysis: past and present. In: Proceedings of the 6th international conference on pattern recognition. IEEE Computer Society Press, p 1152–68

Coulborn RM, Panunzi I, Spijker S, Brant WE, Duran LT, Kosack CS, Murowa MM (2012) Feasibility of using teleradiology to improve tuberculosis screening and case management in a district hospital in Malawi. Bull World Health Organ 90:705–711

Dawson R, Masuka P, Edwards DJ, Bateman ED, Bekker L, Wood R, Lawn SD (2010) Chest radiograph reading and recording system : evaluation for tuberculosis screening in patients with advanced HIV. 14:52–58

Day JH, Charalambous S, Fielding KL, Hayes RJ, Churchyard GJ, Grant AD (2006) Screening for tuberculosis prior to isoniazid preventive therapy among HIV-infected gold miners in South Africa. Int J Tuberc Lung Dis 10:523–529

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics Sep:837–845

Denkinger CM, Nicolau I, Ramsay A, Chedore P, Pai M (2013) Are peripheral microscopy centres ready for next generation molecular tuberculosis diagnostics? Eur Respir J 42:544–7

Dheda K, Barry CE, Maartens G (2015) Tuberculosis. Lancet 6736:1–17

Dye C, Floyd K (2006) Tuberculosis. In: Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A, Musgrove P (eds) Disease Control Priorities in Developing Countries (2nd Edition). The World Bank & Oxford University Press, Washington DC, New York

Feinstein AR, Cicchetti D V. (1990) High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol 43:543–549

Field N, Lim M, Murray J, Dowdeswell RJ, Glynn JR, Sonnenberg P (2014) Timing, rates, and causes of death in a large South African tuberculosis programme. BMC

Infect Dis 14:679

FIND (2015) Technology Landscape Report 2015 Digital Radiology Solutions for TB Diagnostics in Low- and Middle-income Countries.

Fourie PB, Gatner EMS, Glatthaar E, Kleeberg HH (1980) Follow-up tuberculosis prevalence survey of Transkei. Tubercle 61:71–79

GBD 2013 Mortality and Causes of Death Collaborators (2014) Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 385:117–71

Ginneken B van, Hogeweg L, Maduskar P, Peters-Bax L, Dawson R, Dheda K, Ayles H, Melendez J, Sanchez C (2012) Performance of inexperienced and experienced observers in detection of active tuberculosis on digital chest radiographs with and without the use of computer-aided diagnosis. In: at: Annual Meeting of the Radiological Society of North America, 2012.

Ginneken B van, Schaefer-Prokop CM, Prokop M (2011) Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261:719–32

GlobalDiagnostiX (2015) GlobalDiagnostiX. URL http://www.globaldiagnostix.org/en (Accessed Dec 07, 2015)

Graham S, Das GK, Hidvegi RJ, Hanson R, Kosiuk J, Al ZK, Menzies D (2002) Chest radiograph abnormalities associated with tuberculosis: reproducibility and yield of active cases. Int J Tuberc Lung Dis 6:137–42

Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, Holland SM, Horsburgh R, Huitt G, Iademarco MF, Iseman M, Olivier K, Ruoss S, Reyn CF Von, Wallace RJ, Winthrop K (2007) An official ATS/IDSA statement: Diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med 175:367–416

Henriques T, Antunes L C-SC (2013) obs.agree: An R package to assess agreement between observers.

Hogeweg L (2013) Automatic detection of tuberculosis in chest radiographs. Radboud University Nijmegen Medical Center

Hogeweg L, Mol C, Jong PA de, Dawson R, Ayles H, Ginneken B van (2010) Fusion of local and global detection systems to detect tuberculosis in chest radiographs. Med Image Comput Comput Assist Interv 13:650–7

Hogeweg L, Story A, Hayward A, Aldridge R, Abubakar I, Maduskar P, Ginneken B van (2011) Computer-aided detection of tuberculosis among high risk groups: potential for automated triage. In: at: Annual Meeting of the Radiological Society of North America 2011.

Hoog AH van't, Langendam MW, Mitchell E, Cobelens FG, Sinclair D, Leeflang MMG, Lonnroth K (2014) Symptom- and chest-radiography screening for active pulmonary tuberculosis in HIV-negative adults and adults with unknown HIV status. Cochrane Database Syst Rev:66

Hoog AH van't, Meme HK, Deutekom H Van, Mithika AM, Olunga C, Onyino F, Borgdorff MW (2011a) High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. Int J Tuberc Lung Dis 15:1308–1314

Hoog AH van't, Meme HK, Deutekom H Van, Mithika AM, Olunga C, Onyino F, Borgdorff MW (2011b) High sensitivity of chest radiograph reading by clinical officers. Int J Tuberc Lung Dis 15:1308–1314

Hoog AH van't, Meme H, Laserson K, Agaya J, Muchiri B, Githui W, Odeny L, Marston B, Borgdorff M (2012) Screening strategies for tuberculosis prevalence surveys: the value of chest radiography and symptoms. PLoS One 7:e38691

International Health Partners US (2015) Children's Hospital at Zinga, Tanzania. Radiology Unit. URL http://www.ihptz.org/files/x-ray.pdf (Accessed Jan 07, 2016)

International Labour Organization (2011) Guidelines for the use of the ILO International Classification of Radiographs of Pneumoconioses, Revised edition 2011.

Jaeger S, Candemir S, Antani S, Wáng YJ, Lu P, Thoma G (2014) Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 4:475–477

Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Zhiyun Xue, Palaniappan K, Singh RK, Antani S, Thoma G, Yi-Xiang Wang, Pu-Xuan Lu, McDonald CJ (2014) Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 33:233–45

Jaeger S, Karargyris A, Candemir S, Siegelman J, Folio L, Antani S, Thoma G (2013) Automatic screening for tuberculosis in chest radiographs: a survey. Quant Imaging Med Surg 3:89–99

Khan A, Zaidi A, Philipsen R, Khowaja S, Ginneken B van, Khan A, Dossul T (2014) Detection of Chest X-ray abnormalities and tuberculosis using computer-aided detection vs interpretation by radiologists and a clinical officer. In: 45th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union). Barcelona, Spain

Kik S V., Denkinger CM, Chedore P, Pai M (2014) Replacing smear microscopy for the diagnosis of tuberculosis: what is the market potential? Eur Respir J 43:1793–1796

Kim YK, Hahn S, Uh Y, Im DJ, Lim YL, Choi HK, Kim HY (2014) Comparable characteristics of tuberculous and non-tuberculous mycobacterial cavitary lung diseases. Int J Tuberc Lung Dis 18:725–729

Kim JY, Shakow A, Castro A, Vande C, Farmer P (2002) Tuberculosis control. In: Smith R, Beaglehole R, Woodward D, Drager N (eds) Global Public Goods for Health. Oxford University Press

Koeslag A, Jager G de (2001) Computer Aided Diagnosis of Miliary Tuberculosis. Proc Pattern Recognit Assoc South Africa

Koppaka R, Bock N (2004) 12. How reliable is chest radiography? In: Toman's Tuberculosis: Case Detection, Treatment , and Monitoring- Questions and Answers.p 51–60

Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106

Kroidl I, Clowes P, Reither K, Mtafya B, Rojas-Ponce G, Ntinginya EN, Kalomo M, Minja LT, Kowuor D, Saathoff E, Kroidl A, Heinrich N, Maboko L, Bates M, O’Grady J, Zumla A, Hoelscher M, Rachow A (2015) Performance of urine lipoarabinomannan assays for paediatric tuberculosis in Tanzania. Eur Respir J 46:761–70

Kundel HL, Polansky M (2003) Measurement of observer agreement. Radiology 228:303–308

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Lange C, Mori T (2010) Advances in the diagnosis of tuberculosis. Respirology 15:220–240

Leibstein JM, Nel AL (2011) Detecting tuberculosis in chest radiographs using image processing techniques. In: Poster presented at the Southern Africa Telecommunication Networks and Applications Conference 2011.

Lemon J (2006) Plotrix: a package in the red light district of R. R-News 6:8–12

Leth F van (2013) First Tuberculosis Prevalence Survey in the United Republic of Tanzania. Primary Analysis. Final Report.

Lewis JJ, Charalambous S, Day JH, Fielding KL, Grant AD, Hayes RJ, Corbett EL, Churchyard GJ (2009) HIV infection does not affect active case finding of tuberculosis in South African gold miners. Am J Respir Crit Care Med 180:1271–8

Lu C, Liu Q, Sarma A, Fitzpatrick C, Falzon D, Mitnick CD (2013) A Systematic Review of Reported Cost for Smear and Culture Tests during Multidrug-Resistant Tuberculosis Treatment. PLoS One 8:e56074

Maduskar P, Adetifa IMO, Hombergh J van den, Leroy-Terquem E, Fasan-Odunsi A, Sanchez C, D’Alessandro U, Ginneken B van (2015) Computerized Reading of Chest Radiographs in The Gambia National Tuberculosis Prevalence Survey: Retrospective Comparison with Human Experts. In: 46th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union). Cape Town, South Africa

Maduskar P, Hogeweg L, Philipsen R, Ginneken B van (2013) Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs (CL Novak and S Aylward, Eds.). :867038–867038–6

Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, Ginneken B van (2013) Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. Int J Tuberc Lung Dis 17:1613–20

Maru DS-R, Schwarz R, Jason A, Basu S, Sharma A, Moore C (2010) Turning a blind eye: the mobilization of radiology services in resource-poor regions. Global Health 6:18

Meyer D, Zeileis A, Hornik K (2015) Visualizing Categorical Data. R package version 1.4-1.

Mhimbira FA, Bholla M, Sasamalo M, Mukurasi W, Hella JJ, Jugheli L, Reither K (2015) Detection of Mycobacterium tuberculosis by EasyNAT diagnostic kit in sputum samples from Tanzania. J Clin Microbiol 53:1342–1344

Murray CJL, Ortblad KF, Guinovart C, Lim SS, Wolock TM, Roberts DA, Dansereau EA, Graetz N, Barber RM, Brown JC, Wang H, Duber HC, Naghavi M, Dicker D, Dandona L, Salomon JA, Heuton KR, Foreman K, Phillips DE, Fleming TD, Flaxman AD, Phillips BK, Johnson EK, Coggeshall MS, Abd-Allah F, Abera SF, Abraham JP, Abubakar I, Abu-Raddad LJ, Abu-Rmeileh NM, Achoki T, Adeyemo AO, Adou AK, Adsuar JC, Agardh EE, Akena D, Kahbouri MJ Al, Alasfoor D, Albittar MI, Alcalá-Cerra G, Alegretti MA, Alemu ZA, Alfonso-Cristancho R, Alhabib S, Ali R, Alla F, Allen PJ, Alsharif U, Alvarez E, Alvis-Guzman N, Amankwaa AA, Amare AT, Amini H, Ammar W, Anderson BO, Antonio CAT, Anwari P, Ärnlöv J, Arsenijevic VSA, Artaman A, Asghar RJ, Assadi R, Atkins LS, Badawi A, Balakrishnan K, Banerjee A, Basu S, Beardsley J, Bekele T, Bell ML, Bernabe E, Beyene TJ, Bhala N, Bhalla A, Bhutta ZA, Abdulhak A Bin, Binagwaho A, Blore JD, Basara BB, Bose D, Brainin M, Breitborde N, Castañeda-Orjuela CA, Catalá-López F, Chadha VK, Chang J-C, Chiang PP-C, Chuang T-W, Colomar M, Cooper LT, Cooper C, Courville KJ, Cowie BC, Criqui MH, Dandona R, Dayama A, Leo D De, Degenhardt L, Pozo-Cruz B Del, Deribe K, Jarlais DC Des, Dessalegn M, Dharmaratne SD, Dilmen U, Ding EL, Driscoll TR, Durrani AM, Ellenbogen RG, Ermakov SP, Esteghamati A, Faraon EJA, Farzadfar F, Fereshtehnejad S-M, Fijabi DO, Forouzanfar MH, Fra.Paleo U, Gaffikin L, Gamkrelidze A, Gankpé FG, Geleijnse JM, Gessner BD, Gibney KB, Ginawi IAM, Glaser EL, Gona P, Goto A, Gouda HN, Gugnani HC, Gupta R, Gupta R, Hafezi-Nejad N, Hamadeh RR, Hammami M, Hankey GJ, Harb HL, Haro JM, Havmoeller R, Hay SI, Hedayati MT, Pi IBH, Hoek HW, Hornberger JC, Hosgood HD, Hotez PJ, Hoy DG, Huang JJ, Iburg KM, Idrisov BT, Innos K, Jacobsen KH, Jeemon P, Jensen PN, Jha V, Jiang G, Jonas JB, Juel K, Kan H, Kankindi I, Karam NE, Karch A, Karema CK, Kaul A, Kawakami N, Kazi DS, Kemp AH, Kengne AP, Keren A, Kereselidze M, Khader YS, Khalifa SEAH, Khan EA, Khang Y-H, Khonelidze I, Kinfu Y, Kinge JM, Knibbs L, Kokubo Y, Kosen S, Defo BK, Kulkarni VS, Kulkarni C, Kumar K, Kumar RB, Kumar GA, Kwan GF, Lai T, Balaji AL, Lam H, Lan Q, Lansingh VC, Larson HJ, Larsson A, Lee J-T, Leigh J, Leinsalu M, Leung R, Li Y, Li Y, Lima GMF De, Lin H-H, Lipshultz SE, Liu S, Liu Y, Lloyd BK, Lotufo PA, Machado VMP, Maclachlan JH, Magis-Rodriguez C, Majdan M, Mapoma CC, Marcenes W, Marzan MB, Masci JR, Mashal MT, Mason-Jones AJ, Mayosi BM, Mazorodze TT, Mckay AC, Meaney PA, Mehndiratta MM, Mejia-Rodriguez F, Melaku YA, Memish ZA, Mendoza W, Miller TR, Mills EJ, Mohammad KA, Mokdad AH, Mola GL, Monasta L, Montico M, Moore AR, Mori R, Moturi WN, Mukaigawara M, Murthy KS, Naheed A, Naidoo KS, Naldi L, Nangia V, Narayan KMV, Nash D, Nejjari C, Nelson RG, Neupane SP, Newton CR, Ng M, Nisar MI, Nolte S, Norheim OF, Nowaseb V, Nyakarahuka L, Oh I-H, Ohkubo T, Olusanya BO, Omer SB, Opio JN, Orisakwe OE, Pandian JD, Papachristou C, Caicedo AJP, Patten SB, Paul VK, Pavlin BI, Pearce N, Pereira DM, Pervaiz A, Pesudovs K, Petzold M, Pourmalek F, Qato D, Quezada AD, Quistberg DA, Rafay A, Rahimi K, Rahimi-Movaghar V, Rahman SU, Raju M, Rana SM, Razavi H, Reilly RQ, Remuzzi G, Richardus JH, Ronfani L, Roy N, Sabin N, Saeedi MY, Sahraian MA, Samonte GMJ, Sawhney M, Schneider IJC, Schwebel DC, Seedat S, Sepanlou SG, Servan-Mori EE, Sheikhbahaei S, Shibuya K, Shin HH, Shiue I, Shivakoti R, Sigfusdottir ID, Silberberg DH, Silva AP, Simard EP, Singh JA, Skirbekk V, Sliwa K, Soneji S, Soshnikov SS, Sreeramareddy CT, Stathopoulou VK, Stroumpoulis K, Swaminathan S, Sykes BL, Tabb KM, Talongwa RT, Tenkorang EY, Terkawi AS, Thomson AJ, Thorne-Lyman AL, Towbin JA, Traebert J, Tran BX, Dimbuene ZT, Tsilimbaris M, Uchendu US, Ukwaja KN, Uzun SB, Vallely AJ, Vasankari TJ, Venketasubramanian N, Violante FS, Vlassov VV, Vollset SE, Waller S, Wallin MT, Wang L, Wang X, Wang Y, Weichenthal S, Weiderpass E, Weintraub RG, Westerman R, White RA, Wilkinson JD, Williams TN, Woldeyohannes SM, Wong JQ, Xu G, Yang YC, Yano Y, Yentur GK, Yip P, Yonemoto N, Yoon S-J, Younis M, Yu C, Jin KY, Sayed Zaki M El, Zhao Y, Zheng Y, Zhou M, Zhu J, Zou XN, Lopez AD, Vos T (2014) Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 384:1005–1070

Muto H, Tani Y, Suzuki S, Yokooka Y, Abe T, Sase Y, Terashita T, Ogasawara K (2011) Filmless versus film-based systems in radiographic examination costs: an activity-based costing method. BMC Health Serv Res 11:246

Muyoyeta M, Maduskar P, Moyo M, Kasese N, Milimo D, Spooner R, Kapata N, Hogeweg L, Ginneken B van, Ayles H (2014) The Sensitivity and Specificity of Using a Computer Aided Diagnosis Program for Automatically Scoring Chest X-Rays of Presumptive TB Patients Compared with Xpert MTB/RIF in Lusaka Zambia. PLoS One 9:e93757

MVIP Software+Consulting GmbH, Verhey M DigiPortXCAD © -Computer Aided Diagnosis. URL http://cms.mvip.de/projekte/ (Accessed Jan 01, 2016)

Niemz A (Keck GI, Boyle DS (Program for AT in H (2012) Nucleic acid testing for tuberculosis at the point-of-care in high-buden countries. Expert Rev Mol Diagnostics 12:687–701

NTLP Tanzania (2013) Manual for the Management of Tuberculosis and Leprosy. National Tuberculosis and Leprosy Programme, Ministry of Health and Social Welfare, Dar es Salaam

Odone A, Amadasi S, White RG, Cohen T, Grant AD, Houben RMGJ (2014) The Impact of Antiretroviral Therapy on Mortality in HIV Positive People during Tuberculosis Treatment: A Systematic Review and Meta-Analysis. PLoS One 9:e112017

Padmapriyadarsini C, Tripathy S, Sekar L, Bhavani PK, Gaikwad N, Annadurai S, Narendran G, Selvakumar N, Risbud AR, Sheta D, Rajasekaran S, Thomas A, Wares F, Swaminathan S (2013) Evaluation of a diagnostic algorithm for sputum smear-negative pulmonary tuberculosis in HIV-infected adults. J Acquir Immune Defic Syndr 63:331–8

Pan American Health Organization (2012) World Radiography Day: Two-Thirds of the World’s Population has no Access to Diagnostic Imaging. URL http://www.paho.org/hq/index.php?option=com_content&view=article&id=7410%3A2012-dia-radiografia-dos-tercios-poblacion-mundial-no-tiene-acceso-diagnostico-imagen&catid=740%3Anews-press-releases&Itemid=1926&lang=en (Accessed Dec 12, 2015)

Pande T, Pai M, Khan F, Denkinger C (2015) Use of chest radiography in the 22 highest tuberculosis burden countries. Eur Respir J 13:1–4

Pantoja A, Fitzpatrick C, Vassall A, Weyer K, Floyd K (2013) Xpert MTB/RIF for diagnosis of tuberculosis and drug-resistant tuberculosis: a cost and affordability analysis. Eur Respir J 42:708–720

Perry L, Malkin R (2011) Effectiveness of medical equipment donations to improve health systems: how much medical equipment is broken in the developing world? Med Biol Eng Comput 49:719–22

Petrone L, Cannas A, Aloi F, Nsubuga M, Sserumkuma J, Nazziwa RA, Jugheli L, Lukindo T, Girardi E, Reither K, Goletti D (2015) Blood or Urine IP-10 Cannot Discriminate between Active Tuberculosis and Respiratory Diseases Different from Tuberculosis in Children. Biomed Res Int 2015:1–11

Philipsen RHHM, Maduskar P, Hogeweg L, Ginneken B van (2013) Normalization of chest radiographs. Proc SPIE 8670:86700G–86700G–6

Philipsen RHHM, Sanchez CI, Maduskar P, Melendez J, Ginneken B van, Lew W (2015) Objective Computerized Chest Radiography Screening to Detect Tuberculosis in the Philippines. In: 46th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union). Cape Town, South Africa

Philipsen RHHM, Sánchez CI, Maduskar P, Melendez J, Peters-Bax L, Peter JG, Dawson R, Theron G, Dheda K, Ginneken B van (2015) Automated chest-radiography as a triage for Xpert testing in resource-constrained settings: a prospective study of diagnostic accuracy and costs. Sci Rep 5:12215

Pinto LM, Dheda K, Theron G, Allwood B, Calligaro G, Zyl-Smit R van, Peter J, Schwartzman K, Menzies D, Bateman E, Pai M, Dawson R (2013) Development of a simple reliable radiographic scoring system to aid the diagnosis of pulmonary tuberculosis. PLoS One 8:e54235

Portevin D, Moukambi F, Clowes P, Bauer A, Chachage M, Ntinginya NE, Mfinanga E, Said K, Haraka F, Rachow A, Saathoff E, Mpina M, Jugheli L, Lwilla F, Marais BJ, Hoelscher M, Daubenberger C, Reither K, Geldmacher C (2014) Assessment of the novel T-cell activation marker-tuberculosis assay for diagnosis of active tuberculosis in children: A prospective proof-of-concept study. Lancet Infect Dis 14:931–938

R Core Team (2015) R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ralph AP, Ardian M, Wiguna A, Maguire GP, Becker NG, Drogumuller G, Wilks MJ, Waramori G, Tjitra E, Sandjaja, Kenagalem E, Pontororing GJ, Anstey NM, Kelly PM (2010) A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. Thorax 65:863–9

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

Reither K, Jugheli L, Glass TR, Sasamalo M, Mhimbira FA, Weetjens BJ, Cox C, Edwards TL, Mulder C, Beyene NW, Mahoney A (2015) Evaluation of giant African pouched rats for detection of pulmonary tuberculosis in patients from a high-endemic setting. PLoS One 10:1–13

Reither K, Manyama C, Clowes P, Rachow A, Mapamba D, Steiner A, Ross A, Mfinanga E, Sasamalo M, Nsubuga M, Aloi F, Cirillo D, Jugheli L, Lwilla F (2014) Xpert MTB/RIF assay for diagnosis of pulmonary tuberculosis in children: A prospective, multi-centre evaluation. J Infect 70:392–399

Riley RL, Mills CC, O’Grady F, Sultan LU, Wittstadt F, Shivpuri DN (1962) Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. Am Rev Respir Dis 85:511–25

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77

Samulski M, Hupse R, Boetes C, Mus RDM, Heeten GJ den, Karssemeijer N (2010) Using computer-aided detection in mammography as a decision support. Eur Radiol 20:2323–30

Shah S, Demissie M, Lambert L, Ahmed J, Leulseged S, Kebede T, Melaku Z, Mengistu Y, Lemma E, Wells CD, Wuhib T, Nelson LJ (2009) Intensified tuberculosis case finding among HIV-Infected persons from a voluntary counseling and testing center in Addis Ababa, Ethiopia. J Acquir Immune Defic Syndr 50:537–45

Siddiqi K, Lambert M-L, Walley J (2003) Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. Lancet Infect Dis 3:288–96

Sreeramareddy, Panduru K, Menten J, Ende J Van den (2009) Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature. BMC Infect Dis 9:91

Steiner A, Mangu C, Hombergh J van den, Deutekom H van, Ginneken B van, Clowes P, Mhimbira F, Mfinanga S, Rachow A, Reither K, Hoelscher M (2015) Screening for pulmonary tuberculosis in a Tanzanian prison and computer-aided interpretation of chest X-rays. Public Heal Action 5:249–254

Stevenson M, Nunes T, Sanchez J, Thornton R, Reiczigel J, Robison-Cox J, Sebastiani P, Solymos P (2013) epiR - An R package for the analysis of epidemiological data.

Storla DG, Yimer S, Bjune GA (2008) A systematic review of delay in the diagnosis and treatment of tuberculosis. BMC Public Health 8:15

Story A, Aldridge RW, Abubakar I, Stagg HR, Lipman M, Watson JM, Hayward a C (2012) Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study. Int J Tuberc Lung Dis 16:1461–7

Swindells S, Komarow L, Tripathy S, Cain KP, MacGregor RR, Achkar JM, Gupta A, Veloso VG, Asmelash A, Omoz-Oarhe AE, Gengiah S, Lalloo U, Allen R, Shiboski C, Andersen J, Qasba SS, Katzenstein DK (2013) Screening for pulmonary tuberculosis in HIV-infected individuals: AIDS Clinical Trials Group Protocol A5253. Int J Tuberc Lung Dis 17:532–539

Tanimura T, Jaramillo E, Weil D, Raviglione M, Lonnroth K (2014) Financial burden for tuberculosis patients in low- and middle-income countries: a systematic review. Eur Respir J 43:1763–1775

Tanzanian Ministry of Health and Social Welfare, University) I (Columbia, International Support for Pulmonology PF (2010) Chest X-ray interpretation in TB/HIV setting training course [Presentation].

The Radiological Society of North America (2015) Patient Safety - Radiation Dose in X-Ray and CT Exams. URL http://www.radiologyinfo.org/en/info.cfm?pg=safety-xray (Accessed Jan 06, 2016)

Tiemersma EW, Werf MJ van der, Borgdorff MW, Williams BG, Nagelkerke NJD (2011) Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review. PLoS One 6:e17601

Tuberculosis Coalition for Technical Assistance (TBCTA) (2010) Handbook for District Hospitals in Resource Constrained Settings for the Quality Improvement of Chest X-ray Reading in Tuberculosis Suspects.

Ukwaja KN, Modebe O, Igwenyi C, Alobu I (2012) The economic burden of tuberculosis care for patients and households in Africa: a systematic review. Int J Tuberc Lung Dis 16:733–739

Waitt CJ, Joekes EC, Jesudason N, Waitt PI, Goodson P, Likumbo G, Kampondeni S, Faragher EB, Squire SB (2013) The effect of a tuberculosis chest X-ray image reference set on non-expert reader performance. Eur Radiol:1–5

Weyer K, Mirzayev F, Migliori GB, Gemert W Van, D’Ambrosio L, Zignol M, Floyd K, Centis R, Cirillo DM, Tortoli E, Gilpin C, Dieu Iragena J de, Falzon D, Raviglione M (2013) Rapid molecular TB diagnosis: evidence, policy making and global implementation of Xpert MTB/RIF. Eur Respir J 42:252–271

WHO (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. Who/Htn/Tb/2007 379

WHO (2011) WHO | Database of national HIV and TB guidelines. URL http://www.who.int/hiv/pub/national_guidelines/en/ (Accessed Feb 04, 2016)

WHO (2013a) Shorter treatment regimens for multidrug-resistant tuberculosis (MDR-TB).

WHO (2013b) Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children: Policy Update.

WHO (2013c) Systematic screening for active tuberculosis: Principles and Recommendations. Who/Htm/Tb/201304:1–123

WHO (2014a) TB: Reach the 3 Million. Geneva, Switzerland

WHO (2014b) Country profile Tanzania. In: Global atlas of medical devices, 2014 updat. Geneva

WHO (2014c) High-priority target product profiles for new tuberculosis diagnostics : report of a consensus meeting.

WHO (2015a) Global tuberculosis report 2015. Geneva, Switzerland

WHO (2015b) Tuberculosis Fact sheet No. 104.

WHO (2015c) Implementing tuberculosis diagnostics: A policy framework. : 39

WHO (2015d) The END TB strategy. WHO/HTM/TB

Wickham H (2007) Reshaping Data with the reshape Package. J Stat Softw 21:1–20

Wickham H (2009) ggplot2: elegant graphics for data analysis.

Xu T, Cheng I, Mandal M (2011) Automated cavity detection of infectious pulmonary tuberculosis in chest radiographs. Conf Proc IEEE Eng Med Biol Soc 2011:5178–81

Zaidi A, Khalid N, Philipsen R, Ginneken B van, Khowaja S, Khan A (2014) Symptomatic screening and computer-aided radiography for active-case finding of tuberculosis: a prediction model for TB case detection. In: 45th World Conference on Lung Health of the International Union against Tuberculosis and Lung Disease (The Union). Barcelona, Spain

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 10:1123–6

Zennaro F, Oliveira Gomes JA, Casalino A, Lonardi M, Starc M, Paoletti P, Gobbo D, Giusto C, Not T, Lazzerini M (2013) Digital Radiology to Improve the Quality of Care in Countries with Limited Resources: A Feasibility Study from Angola. PLoS One 8:e73939

Zumla A, Nahid P, Cole ST (2013) Advances in the development of new tuberculosis drugs and treatment regimens. Nat Rev Drug Discov 12:388–404

Erklärung zum Eigenanteil

<u>Studienkonzeption:</u>

Die Idee zu dieser Studie hatte Dr. Klaus Reither vom Swiss Tropical and Public Health Institute (Swiss TPH), Basel, Schweiz. In enger Absprache verfasste die Doktorandin Marianne Breuninger hierfür selbstständig einen Studienentwurf. Dr. Levan Jugheli, Prof. Dirk Wagner, Dr. Jan van den Hombergh und Dr. Fred Lwilla standen beratend zur Seite.

<u>Auswertung der Daten:</u>

Für die Auswertung der Daten erstellte Marianne Breuninger einen Statistischen Analyseplan. Diesen konnte sie mit Amanda Ross, Statistikerin am Swiss TPH, besprechen. Marianne Breuninger führte die statistische Auswertung der Daten selbstständig durch. Bei Fragen konnte sie sich jederzeit an Amanda Ross wenden. Dr. Klaus Reither und Dr. Levan Jugheli standen unterstützend zur Seite.

<u>Datenrecherche:</u>

Marianne Breuninger extrahierte die Röntgenbilder und die zugehörigen klinischen Patientendaten aus den TB Cohort und TB CHILD Datenbanken. Bei der Extraktion der klinischen Daten unterstützten sie Dr. Levan Jugheli, Dr. Andreas Steiner, Dr. Jerry Hella und Dr. Francis Mhimbira.

Die klinischen Daten und die Röntgenbilder wurden von Dr. Klaus Reither, Leiter der TB Cohort und TB CHILD Studien in Bagamoyo zur Verfügung gestellt. Die CAD4TB-Ergebnisse stammten von Rick Philipsen und Prof. Bram van Ginneken von der Diagnostic Image Analysis Group des Radboud University Medical Centers, Nijmegen, Niederlande. Die Röntgenbilder wurden eigens für diese Studie von Dr. Jan van den Hombergh, Dr. Jaffer Dharsee und Mwinyikambi Salum interpretiert.

<u>Betreuung der Arbeit:</u>

Die Betreuung erfolgte durch Dr. Klaus Reither (Swiss TPH) und Prof. Dirk Wagner (Abteilung für Infektiologie, Universitätsklinik Freiburg).

Schreiben der Publikation:

Die Publikation (Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa, PLoS One (2014)) wurde von Marianne Breuninger, Dr. Klaus Reither und Amanda Ross verfasst. Eine kritische Durchsicht des Manuskripts erfolgte durch Prof. Bram van Ginneken, Rick Philipsen, Dr. Francis Mhimbira, Dr. Jerry Hella, Dr. Fred Lwilla, Dr. Jan van den Hombergh, Prof. Dirk Wagner und Dr. Levan Jugheli.