The diagnostic accuracy of artificial intelligence based computer programs for analyzing chest x-rays to detect pulmonary tuberculosis: a systematic review

Miriam Harris

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec

2018-12-16

Report of thesis carried out as a requirement of the M.Sc. Thesis Program in Biostatistics, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

DEDICATION

This thesis is dedicated to my parents, Beth Henning & Stewart Harris

TABLE OF CONTENTS

ABSTRACT

Background: Chest radiography has been widely used for tuberculosis (TB) detection in developed countries for over a century (1) but uptake of chest x-rays (CXR) in high TB burden, resource-constrained countries has been limited (2, 3). Obstacles to greater CXR usage in such settings include the high costs of radiographic films and the paucity of professionals to interpret images (4). Artificial intelligence (AI) - based software for identification of radiologic abnormalities (*computer-aided detection*, or CAD) on CXRs could address these problems and potentially facilitate broader use of CXR for the detection of PTB.

Methods: We undertook a systematic review of the diagnostic accuracy of using CAD to detect abnormalities compatibles with pulmonary TB (PTB) on CXR. We searched four databases for articles published between January 2005-November 2017. We summarized data on CAD type, study design, and measures of diagnostic accuracy. We assessed risk of bias with the QUADAS-2 method. Meta-analyses were not performed due to differences in study design. To summarize the literature, we divided the included reports into development studies and clinical studies.

Results: We included 43 of the 3978 articles reviewed, 11 clinical studies, and 32 development studies. Development studies were more likely to use CXR databases that had greater potential for bias as compared to clinical studies. Areas under the receiver operating characteristic curve (AUC) were significantly higher: in development studies (AUC: 0.87 [0.81-0.90]) versus clinical studies (0.68 [0.65-0.75]; p-value 0.004); and with programs using deep learning versus (0.93 [0.90-0.97]) vs machine-learning (0.81 [0.69-0.88]; p=0.001) software. The sensitivity of clinical studies was higher when using nucleic acid amplification testing as the microbiologic reference standard compared to culture.

Conclusion: We conclude that AI-based CAD programs are promising, but much of the work thus far is in the development rather than the clinical phase. This review provides concrete suggestions on what study design elements should be improved.

RESUMÉ

Contexte : Depuis plus d'un siècle, la radiographie pulmonaire (RXP) est largement utilisée pour diagnostiquer la tuberculose (TB) dans les pays développés. Cependant, les pays à forte charge de tuberculose sont limités dans l'utilisation de la RXP et l'accès aux ressources est restreint. Le coût élevé des films radiographiques et le manque de professionnels pour interpréter les radiographies constituent des obstacles à une plus grande utilisation de la RXP dans de tels contextes. -Un logiciel basé sur l'intelligence artificielle (IA) qui détecte des anomalies radiologiques (la détection assistée par ordinateur ou DAO) et qui est adapté à la tuberculose pulmonaire aux RXP - pourrait résoudre ces problèmes et peut-être faciliter une utilisation plus vaste de la RXP dans le diagnostic de la tuberculose pulmonaire (TP).

Méthodes: Nous avons entrepris une revue systématique de l'exactitude du diagnostic effectué par les logiciels basés sur l'IA afin de détecter les anomalies radiologiques (DAO) qui sont compatibles avec la TP par les RXP. Nous avons sondé quatre bases de données pour trouver des articles publiés entre janvier 2005 et novembre 2017. Nous avons résumé les données sur le type de DAO, l'étude clinique et les mesures de l'exactitude du diagnostic. Nous avons évalué le risque de biais avec la méthode QUADAS-2. Aucune méta-analyse n'a été effectuée en raison de différences entre les études cliniques. En résumé, nous avons divisé les rapports inclus en études de développement et en études cliniques.

Résultats: Nous avons inclus 43 des 3978 articles examinés, 11 études cliniques sur la tuberculose CAD4TB et 32 études de développement. Les études de développement étaient plus susceptibles d'utiliser des bases de données provenant de la RXP qui présentaient un plus grand risque de biais que les études cliniques. Les résultats des aires sous la courbe (AUC) de ROC (caractéristique de fonctionnement du récepteur) étaient significativement plus élevés dans les études de développement (AUC : 0,87 [0,81 – 0,90]) que dans les études cliniques (0,68 [0,65 – 0,75]; p= 0,004), ainsi que dans les logiciels qui fonctionnent avec l'apprentissage automatique (0,93 [0,90 – 0,97]) comparativement à ceux qui fonctionnent avec l'apprentissage

approfondi (0,81 [0,69 – 0,88]; *p*=0,001). L'utilisation d'amplification des acides nucléiques comme référence microbiologique dans les études cliniques est une méthode plus sensible comparativement à celles des cultures.

Conclusion : Les programmes de DAO basés sur l'IA sont prometteurs, mais une grande partie du travail effectué jusqu'à présent se situe dans la phase de développement plutôt que dans la phase clinique. Cet examen de la littérature fournit des suggestions réelles sur les éléments de conception des études qui devraient être améliorés.

PREFACE

This thesis aims to systematically review and critically evaluate the literature on computer aided detection (CAD) for PTB on CXRs. This is a manuscript-based thesis written in compliance with the guidelines and specifications detailed by the Faculty of Graduate and Postdoctoral Studies of McGill University.

Chapter 1 reviews the pathophysiology of TB, current epidemiology, and the role of CXR in pulmonary TB diagnosis. Chapter 2 provides an overview of the history of CAD technology, how CAD programs are developed, and how CAD is more specifically applied to CXR reading and PTB identification. Chapter 3 summarizes the aims of the thesis and rational for the selected methodology. Chapter 4 is the manuscript that has been developed and submitted and is currently under review at PLOS One. There may appear to be some repetition in the contents of the manuscript, but that is because it was derived from this thesis. Chapter 5 provides additional details around the limitations and issues with the existing literature. These details required a stand-alone chapter as they could not be succinctly incorporated into a publishable manuscript. Chapter 6 summarizes the findings, reviews the implications of the results, and how they relate to future research avenues.

CONTRIBUTION OF AUTHORS

Miriam Harris (Thesis candidate):

I was responsible for collaborating on the development of the thesis topic and acquiring ethics approval. I was also responsible for the article selection, data collection, data analysis, and writing of this thesis document.

Dr. Faiz Ahmad Khan (Co-Supervisor):

Dr. Ahmad Khan generated the original thesis topic idea and provided insight into the study design, methodology, and statistical plan. Dr. Ahmad Khan acted as a reviewer for article selection, data extraction, and quality review of the included articles. Dr. Ahmad Khan reviewed the thesis document and manuscript.

Dr. Richard Menzies (Supervisor):

Dr. Menzies provided guidance for the thesis approach, writing, and submission. Dr. Menzies reviewed the thesis document and edited the manuscript.

Dr. Amy Qi (Resident):

Dr. Amy Qi acted as a reviewer for article selection and data extraction.

ACKNOWLEDGMENTS

I would like to begin by thanking the members of my thesis committee: my supervisors, Dr. Faiz Ahmad Khan and Richard Menzies.

I am extremely grateful to Dr. James Johnston for putting me in touch with Dr. Faiz Ahmad Khan when I came to McGill from the University of British Columbia for my fellowship in General Internal Medicine. Dr. Ahmad Khan's mentoring and academic support has been invaluable to me during my time at McGill. He has taken time not only to discuss our work together, but also to offer personal career guidance, clinical expertise, and encouragement during times of need. His positive attitude, incredible work ethic, and collegiality represent an inspiring model of a burgeoning clinician-scientist, and I hope I will someday be able to establish a similar career path. This thesis would not have been possible without Dr. Richard Menzies. His great expertise in the study of tuberculosis, and years of academic experience at McGill grounded this thesis work. I am honored to have had the opportunity to work with such a prolific researcher.

I would also like to thank the members of the Respiratory Epidemiology and Clinical Research Unit for creating such a welcoming work environment. In particular, I would like to thank Alix Zerbo and Saeedeh Moayedi Nia for orienting me to the department, and for their help with this thesis work.

I am very grateful to my division of General Internal Medicine for supporting my academic work during my final year residency, and providing funding for my thesis project.

Lastly, I would like to thank my personal supports. I am endlessly appreciative of my parents, Drs. Beth Henning and Stewart Harris, for their love and support throughout my life. My mother is my constant source of inspiration, validation, and encouragement. I consider my father to be my overall academic supervisor, as he is the one I turn to first with questions, as a source of motivation, and for help in my career.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBERVIATIONS/DEFINITIONS

AIDS: acquired immune deficiency syndrome
AFB: acid-fast bacilli
CAD: computer aided detection
CAD4TB: a proprietary software for computer aided detection program, owned by Delft Inc (Veenendaal, Netherlands)
CXR: chest x-ray
EPOC: Effective Practice Organization of Care
HIV: human immunodeficiency virus
IQR: interquartile range
MTB: *mycobacterium tuberculosis*
NAAT: nucleic acid amplification test
PICO: population, intervention, comparator, outcome
PTB: pulmonary tuberculosis
TB: tuberculosis
QUADAS: Quality Assessment of Diagnostic Accuracy Studies
WHO: World Health Organization

Culture: cultivation of Mycobacterium tuberculosis in a growth medium. It is the gold standard for TB diagnosis. Both liquid media and solid media are used. Solid medium requires 4-8 weeks for growth, whereas liquid media is significantly faster (between 10 and 14 days) and may increase the case yield by 10% over solid media (5).

Smear: sputa examined under microscopy for the identification of *Mycobacterium tuberculosis* with either fluorescence microscopy (auramine-rhodamine staining) or Ziehl-Neelsen staining (5). It is inexpensive and identifies the most infectious TB cases; however, it has lower sensitivity to detect TB, especially among people living with HIV.

Xpert MTB/RIF: it is a nucleic acid amplification test (NAAT) for for rapid diagnosis of active TB, as well as resistance to rifampicin (6).

Bacteriologically confirmed active TB: TB disease in which a biological specimen tests positive for *M. Tuberculosis* by demonstration of the organism's DNA via PCR, or of acid-fast bacilli by sputum smear microscopy (acid-fast staining), or by mycobacterial culture (7).

Clinically diagnosed active TB: TB disease diagnosed based on clinical findings (e.g. symptoms and abnormalities seen on CXR) but not fulfilling criteria for bacteriologically confirmed active TB.

Screening use-case: the use of CXRs in the assessment of people belonging to a target group considered at high risk for TB. It does not include use of CXR for evaluation of individuals who seek medical care for respiratory symptoms (see below, Triage use-case) (7).

Triage use-case: use of CXRs to evaluate patients who have sought care with symptoms or signs suggestive of TB (7).

Development study: a CAD study that primarily focused on reporting methods for creating a CAD program for PTB as Development studies. These were often published in engineering, computer science, medical imaging journals, or proceedings from engineering or medical imaging conferences.

Clinical study: a CAD study that primarily focused on the assessment of the accuracy of an already-developed CAD software as clinical studies.

Artificial intelligence: Computer technology with the ability to automate intellectual tasks.

Machine learning: A type of artificial intelligence that relies less on human specification (i.e. defining a set of variables to be included) and instead allows algorithms to decide what variables are important (8).

Deep Learning: is a type of artificial intelligence which attempts to model brain architecture.(9) It uses neural networks, or overlaying models, that emphasize learning increasingly meaningful representations of the data (9).

CHAPTER 1: Introduction

Tuberculosis: pathophysiology & epidemiology

Tuberculosis (TB) remains a major global health problem. It is an infectious disease caused by the bacteria known as *Mycobacterium tuberculosis* (*MTB*) (10). In humans, TB is clinically dichotomized into two forms: latent and active TB; the former occurs after infection when the body's immune system contains the disease, and M. tuberculosis bacilli remain in a dormant (slowly replicating) state that is not associated with progressive tissue destruction . Individuals with latent TB remain asymptomatic but are at risk of developing active TB disease in the future. Active TB is a state of disease, where M. tuberculosis bacilli are replicating and eliciting an immune reaction that together are resulting in tissue destruction (11). It most commonly affects the pulmonary system, but can involve extrapulmonary sites throughout the body (11). This thesis will focus on the most prevalent type of active TB, pulmonary TB (PTB).

TB transmission occurs when someone with active respiratory TB (including the lungs, larynx and upper airways) expels organisms into the air, typically by coughing, and others inhale droplets containing the expelled bacteria (10). Most individuals infected with TB do not develop symptoms. In fact, only about 5-15% of the estimated 1.7 billion people infected with *MTB* develop active TB during their lifetime (10). However, individuals with other co-morbid immune modulating disease, such as human immune deficiency virus (HIV), undernutrition, cancer, and chronic diseases like diabetes or kidney diseases are at much higher risk of developing active TB (10).

Patients are typically diagnosed when they present to health care settings with symptoms suggestive of PTB or through screening programs that target high-risk populations. Confirmatory diagnostic tests for PTB include polymerase chain reaction (PCR) based rapid molecular tests (nucleic acid amplification testing (NAAT)) such as the Xpert MTB/RIF assay

(Cepheid, USA); sputum smear microscopy (acid-fast staining), or mycobacterial culture (10). However, a large portion of TB cases reported to the World Health Organization (WHO) remain clinically diagnosed rather than microbiologically; for instance in 2016, 57% of reported pulmonary cases were diagnosed clinically (10).

The mortality from TB without treatment is very high. Studies predating the availability of effective treatment found that 70% of sputum smear-positive individuals, and 20% of culture-positive/smear-negative individuals died within 10 years of diagnosis (10). Effective medications for the treatment of TB were first developed in the 1940s, but despite access to these, TB remains in the top ten causes of death worldwide, and is the leading cause from a single infectious etiology ahead of HIV/AIDS (10). In 2016, there were an estimated 1.67 million deaths, and only 6.3 million or 61% out of the estimated 10.4 million new cases were reported to national TB programmes (10). Although this is an improvement from 2015, where 6.1 million or 59% of TB cases were diagnosed, TB identification still remains a major cause of the ongoing morbidity and mortality associated with TB infection today (10).

Low and middle income countries (LMIC) suffer from the highest rates of TB, and simultaneously, have the fewest resources for detection and treatment. In countries with the greatest TB burden, radiology, for example, is mainly limited to district level referral centers. Three resource-limited countries account for 76% of the total gap between TB incidence and reported cases, with the top three being India (25%), Indonesia (16%), and Nigeria (8%) (10). The gap between the reported number of new cases compared to the estimated number is due to a combination of factors including underdiagnoses. Closing this gap will require easier access to currently available diagnostic tools and novel diagnostics methods.

Chest x-ray in PTB diagnosis

Chest x-ray (CXR) is a rapid imaging tool that allows lung abnormalities to be visualized. CXR is broadly used to identify conditions of the thoracic cavity, including the lungs, ribs, airways, diaphragm, and heart (12). The role of CXR in TB diagnosis is to identify persons that should

undergo microbiologic testing for TB (12, 13). There are two broad contexts, or use-cases, in which CXR are used for detecting TB. The triage use-case refers to a circumstance where CXRs are being used as part of the diagnostic pathway of someone with symptoms suggestive of PTB (also called "passive case-finding"). The screening use-case refers to the use of CXR in active case finding, when high-risk populations are systematically screened to identify those with active TB (also called “active case-finding). In this use-case patients may not have sought care, and are often asymptomatic.

According to the most recent WHO handbook for *National Tuberculosis Control Programme* and the *International Standards for Tuberculosis Care* 2014 guidelines, in the triage scenario, CXR can be used prior microbiologic testing (7). It may be used as a triage tool to aid in deciding which patients should receive microbiologic investigations (7). Importantly, the WHO emphasized the use of CXR alone to diagnose TB is not recommended given its low specificity, high interobserver variability, poor access to high-quality radiography equipment, lack of expert access for interpretation, and wide-spread use of low-quality radiography (7). However, in resource-constrained settings, the WHO has suggested that CXR along with clinical assessment can be used to triage patients who should be tested with PCR-based rapid molecular tests as a cost reducing strategy (12).

According to the WHO and Canadian guidelines, CXR can be used as a screening tool when available (12, 13). Although CXR is estimated to have only a sensitivity of 70% to 80%, and specificity of 60% to 70%, it is still more sensitive than symptom screen alone, and thus remains useful in the screening setting (13). CXR is particularly useful when used in a screening program that is based in a health care setting or mobile CXR unit (7). While CXR is superior to symptom screen alone, the WHO notes that it can be expensive and logistically challenging to use. Chest radiography has been widely used for TB clinical care and screening in developed countries for over a century (1), but uptake of CXR in high TB burden countries, particularly in resource-constrained settings, has been limited (2, 3).

Two obstacles for greater use of CXR in such settings are the high costs of radiographic films and the paucity of professionals to interpret images (4). The greater affordability of digital radiography technology, and the recent development of computer programs capable of analyzing CXR images to identify PTB compatible abnormalities (computer-aided detection, or CAD), could address these barriers and allow for increased uptake of CXR and improved PTB diagnosis. However, CAD is a new technology and the WHO has called for greater evidence before endorsing its use in TB diagnostic and screening pathways (7).

CHAPTER 2: Computer aided technology for tuberculosis diagnosis

Computer technology has developed the ability to automate intellectual tasks, also known as artificial intelligence (AI). AI has been applied to the analysis of radiologic images to identify abnormalities—referred to as computer-aided detection, or CAD. CAD technologies in radiology, where digital medical images are analyzed quantitatively by computers, date back to the 1960s (14). The term CAD has been operationalized in the literature to mean either that computer analysis is used as a tool to assist human radiologists in decision making, or as a means to replace the human reader altogether. This review will focus on this latter definition of CAD as it applies to CXR reading to detect PTB. The attraction of AI-based CAD for PTB detection highlights the opportunity for CXR utilization in places that currently lack human personnel for image interpretation, thereby addressing the TB detection gap in resource constrained settings.

Two forms of AI have currently been used in image analysis: Machine learning (ML) and Deep learning (DL) (Figure 1). ML is a type of AI analysis that relies less on human specification (i.e. defining a set of variables to be included) and instead allows algorithms to decide what variables are important (8, 9). ML algorithms often involve techniques such as decision trees and association rule learning (8). ML focuses on prediction and classification and can be more amenable than traditional statistical analysis to non-linear, complex, or skewed data (8). Deep learning (DL) is a subset of ML which attempts to model brain architecture. It uses neural networks, or overlaying model layers, that puts an emphasis on learning increasingly meaningful representations of the data (9). DL uses a loss of function to measure the quality of the networks output. It then uses this loss of function data as feedback in order to find the optimally set weights to further optimize the input and achieve a more precise output (Figure 2).

ML or DL software applied to CXR reading are usually trained on a series of images and work by employing serial computational procedures, such as image preprocessing, boundary segmentation, feature extraction, and feature classification to input information into classification algorithms (15, 16). The training is typically completed on a set of CXRs in which abnormal features were identified by human readers. Presented with an unlabeled CXR, the program analyzes the image characteristics and, using the distributions developed with the training datasets determines if PTB is likely or unlikely. This software can then be tested in the clinical setting (Figure 3). A growing body of research on the development and assessment of these new CAD technologies for PTB diagnosis is underway, but their clinical utility is still unclear.

An important question when evaluating CAD software, is whether diagnostic accuracy should be assessed using a set of CXR images that are separate from the training set (i.e. avoid testing accuracy with CXRs that were used for training, or CXRs that were not used for training but that originate from the same subset/study as those with which the program was trained). It is known that using the same set of CXRs for the evaluation will likely lead to an overestimate of the diagnostic accuracy, and thus will have limited generalizability (17). It is also important to consider how the database used for evaluation was designed, in order to know whether the selection of images and the reference standard used could affect the internal and external validity of the assessment of diagnostic accuracy (18, 19). However, such an assessment of the databases that are commonly used to develop and evaluate CAD has never been reported.

Another consideration is that, for clinical-decision making, clinicians use information from CXRs in a categorical manner—that either the image is compatible with PTB or it is not. This information is then used to either test further, for example complete a microbiologic test, or to initiate treatment of PTB. As such, CAD programs should provide an output that can easily be operationalized for categorical decision making, with an established and reproducible level of diagnostic accuracy. A “threshold score” is the abnormality score below which PTB is considered ruled out.

For example, CAD4TB is the only commercially available CAD software developed by Diagnostic Image Analysis Group at Radboud University Medical Center, Nijmegen, The Netherlands. It originally underwent field testing in 2010 (20). It is a machine learning software and the analysis process can be broken down into four steps: (1) the radiographs are processed to standardize the grey scale, resolution, and other features; (2) the software demarcated the anatomical structures such as the lungs, clavicles, ribs, and defined the anatomical orientation of the image; (3) the defined lung fields are then analysed for local texture, shape, and global symmetry; (4) lastly, a global correlation with a typical normal CXR is completed (20). A score from each step is combined to produce an overall score for the image, known as an abnormality score. This score, between 0-100, determines the likelihood of active PTB with higher scores indicating a greater likelihood (20).

This threshold score is currently not preselected by the test developers, in contrast, the test developers advise the user to select the score based on the use-case applied and the population tested. Hence, for CAD programs that report output as a continuous result in the form of a number or score, developers are advised to identify a cut-off, or threshold scores, at which the image can be interpreted as consistent with PTB, and below which PTB can be confidently excluded (16).

Furthermore, when assessing the accuracy of CAD, the context in which the CXR was performed must be considered. The pre-test probability of TB, and the expected prevalence of more advanced or extensive disease, will differ between the two previously described use-cases (triage and screening), thereby affecting the sensitivity of CXR and hence the accuracy of CAD. A study of CAD for PTB identification in the triage use-case may not be generalizable to the identification of PTB in the screening use-case and vice versa.

Chapter 2 Tables and Figures

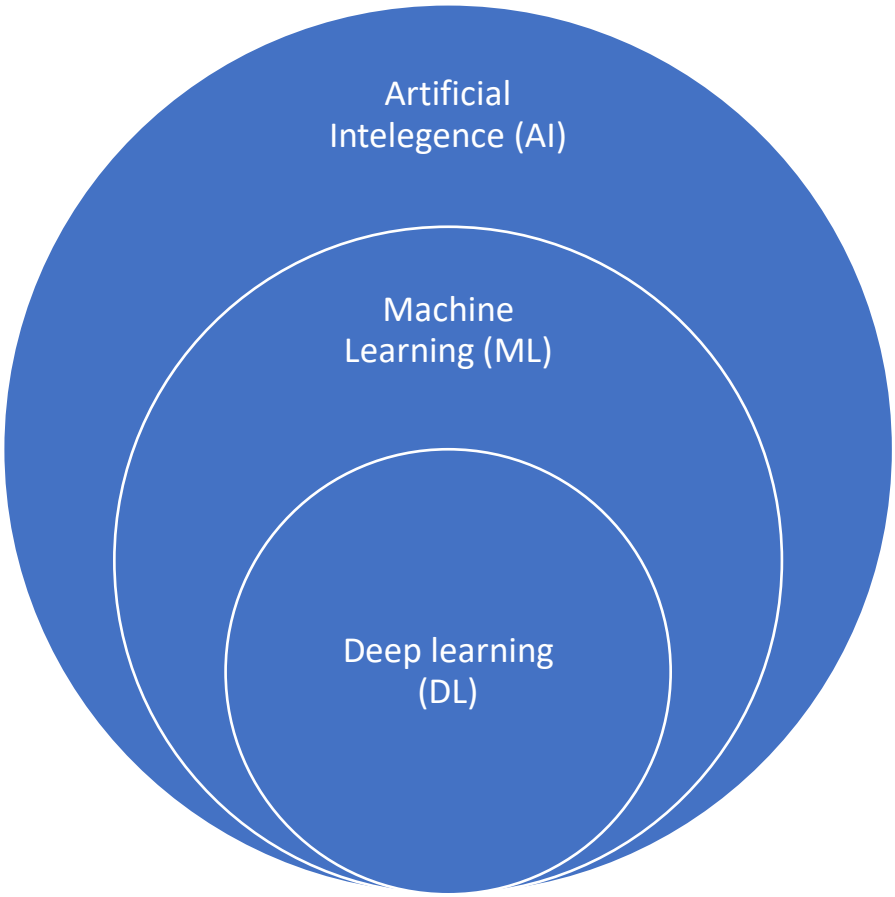

Figure 1. Venn Diagram representation of the hierarchy of terms of artificial intelligence, machine learning and deep learning.

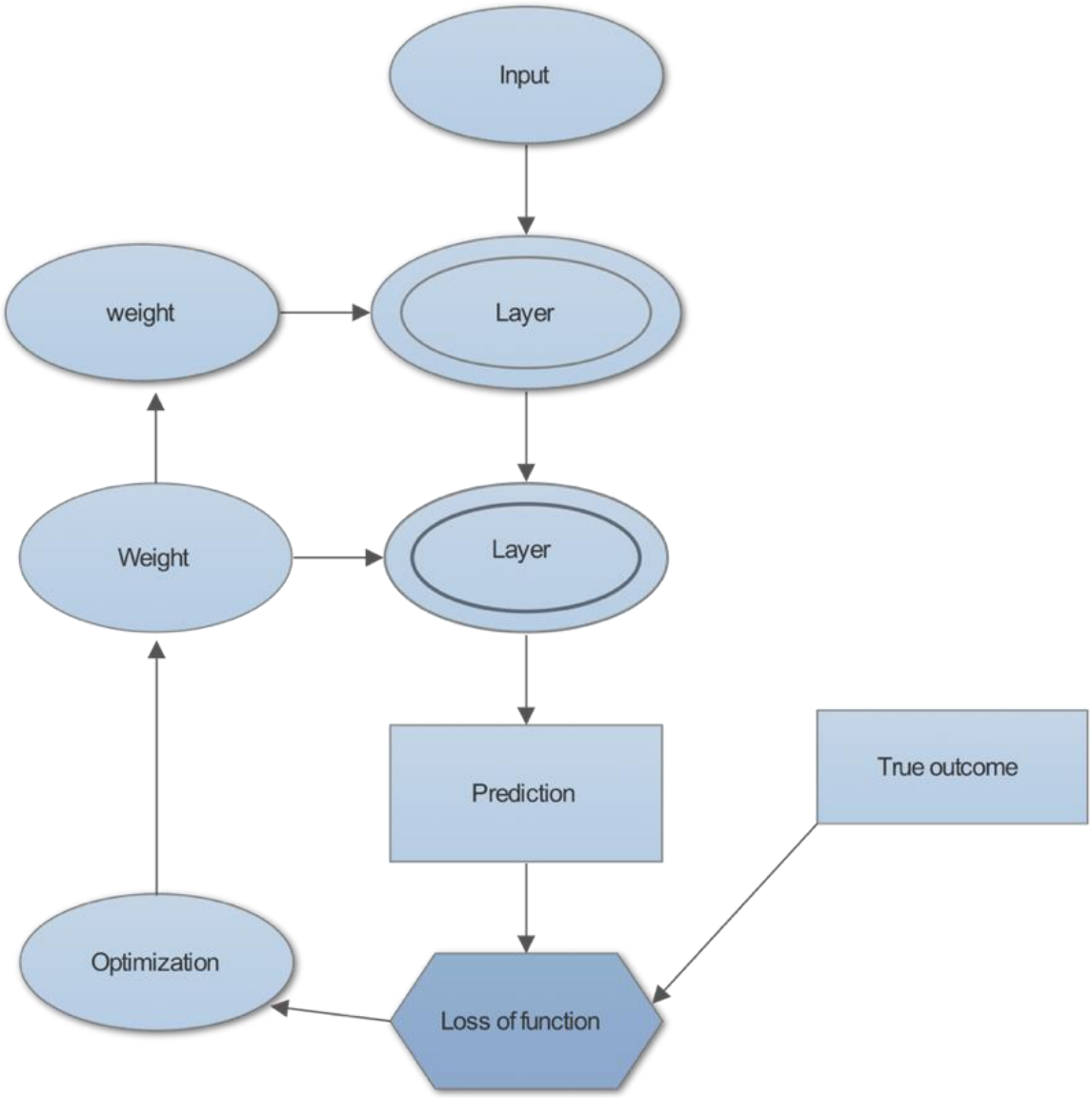

Figure 2. Schematic representation of a deep learning algorithm.

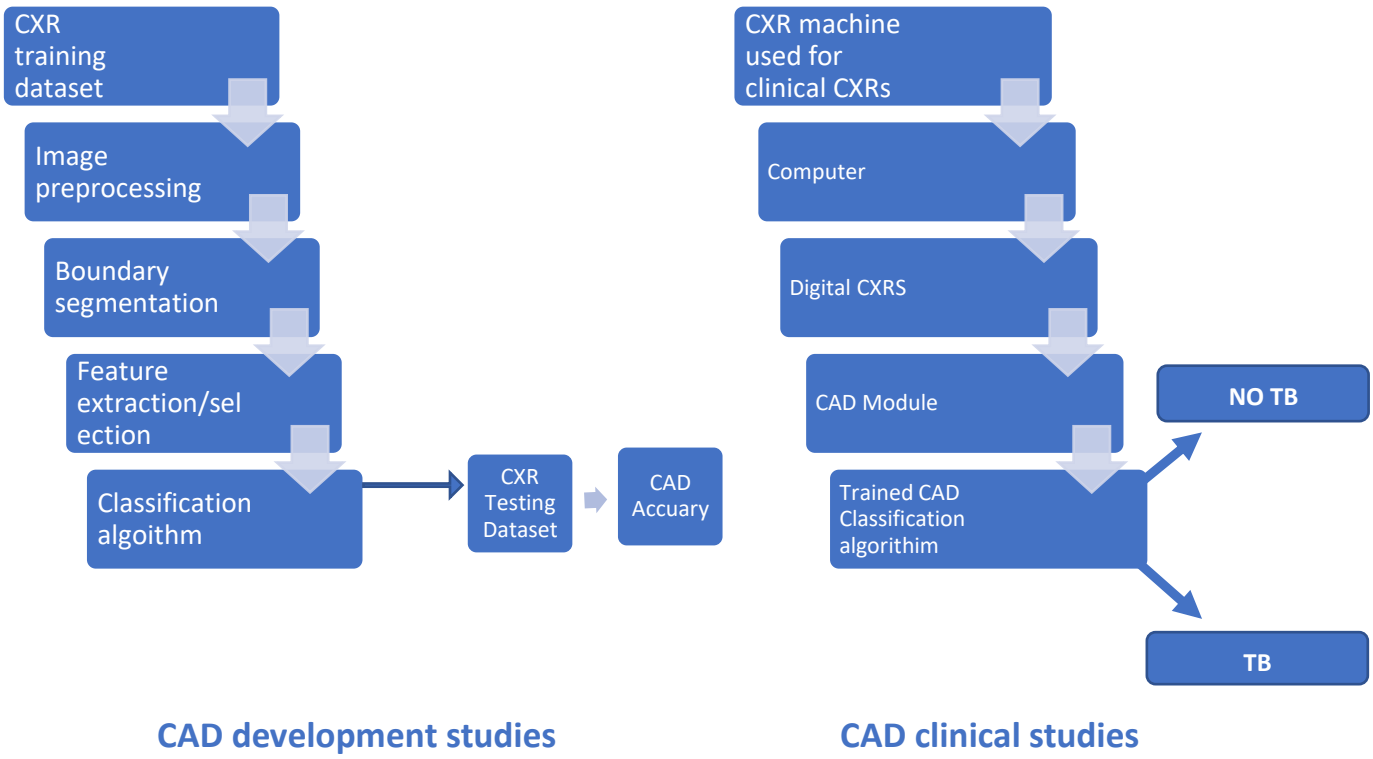

Figure 3. Schematic of representation of CAD development and clinical testing of CXR analysis for TB detection

CHAPTER 3: Thesis approach

In the era of evidence-based medicine, whereby scientific evidence is combined with practitioner expertise to make decisions regarding patient care, systematic reviews are tools used by clinicians and the academic community to identify evidence for a specific research question (21). A systematic review has been defined by the Cochrane Collaboration as "a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review" (22).

In response to the growing body of literature and interest in CAD for PTB diagnosis, this thesis aims to provide a comprehensive summary of the literature on CAD for PTB utilizing a systematic review. A prior systematic review was conducted on the only commercial available CAD software at that time (23). This previous review focused only on CAD4TB. The literature search, conducted in November 2015, identified 5 studies eligible for inclusion. The authors identified a number of limitations that could have resulted in a biased assessment of the diagnostic accuracy and limited the generalizability of the findings. However, because this prior review focused on CAD4TB studies, and did not address non-commercially available studies reporting on the development of CAD programs, the majority of CAD studies were excluded.

Our review also includes non-commercially available CAD studies that focused predominantly on reporting the development methods for CAD programs. Since the publication of the previous review, additional studies of CAD4TB have also been published, and the WHO issued a call for further information (7). This systematic review addresses the following two aims.

AIM 1: To answer the following PICO question:

What is the diagnostic accuracy and clinical utility of CXRs analyzed by CAD for the detection of PTB in the screening and triage scenario?

AIM 2: Which study-level factors affect the reported diagnostic accuracies?

AIM 1

For Aim 1, we completed a systematic review of the literature and addressed diagnostic utility as described by Price et al by looking at 4 aspects: analytical validity, clinical validity, clinical usefulness, and the social context (24). The analytical validity refers to whether the test measures what it is claimed to measure. This is evaluated by the accuracy of the test, and whether it was estimated in a valid manner.

According to Sackett & Haynes, a valid assessment of a diagnostic test includes the following features: an independent, blind comparison of the index test result with a reference standard, assessment within a consecutively enrolled group of individuals, inclusion of missing and indeterminate results, and replication of the studies in other settings (25). To assess the validity of CAD software studies, we applied the widely used 'Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2)', which includes the features outline by Sackett & Haynes (26). We additionally assessed the validity of the databases used. Because many of the studies used the same databases to test and train their software, we summarized the quality by database rather than by CAD study. We adapted the QUADAS-2 approach, as our interest was to assess the databases used for testing CAD performance. Hence, we focused on patient selection and the reference standard used.

The utility of a diagnostic test determines how well the test discriminates between two conditions of interest; for example, disease versus health, two stages of a particular disease, or between two different diseases (27). Diagnostic test outcomes can be dichotomous, ordinal, or continuous. The discriminability of binary tests is quantified by the following measures: accuracy, sensitivity, specificity, positive and negative predictive value, likelihood ratios, area under the receiver operator curve (ROC) (AUC) (27) (Table 1). To assess the accuracy of CAD, we extracted diagnostic accuracy measures when reported from included studies.

The clinical validity addresses whether the test answers the clinical questions being asked. This focuses on how the test should be used i.e. as a screening tool, triage tool, prognostic tool, or for monitoring (24). To assess the clinical validity of CAD software studies, it was important to determine if the database used to test the software was representative of either the triage or screening scenario, as the sensitivity and specificity of test depends on disease prevalence. Clinical studies were analyzed according to their use-case as either screening or triage.

The clinical usefulness, refers to whether the test leads to better outcomes (22). Part of this assessments includes how a test can conceivably be used. Therefore, for CAD, we assessed how the software results were reported and how authors suggested these results be then applied in the clinical context. For clinical decision making, clinicians use information from CXRs in a categorical manner; either the image is compatible with PTB or it is not. As such, we assessed whether CAD programs provided an output that can easily be operationalized for categorical decision making, with an established and reproducible level of diagnostic accuracy. Hence for CAD programs that output used a continuous result in the form of a number or score, we assessed whether developers identified a cut-off, or threshold score, at which the image was interpreted as consistent with PTB, and below which PTB was confidently excluded (16).

Additionally, to assess if CAD leads to better outcomes, we further focused our evaluation on the reference standard authors used to label participants as either having PTB or not. Ultimately, for CAD to be used to diagnosis PTB, microbiologic reference, as the gold standard for diagnosis, should be utilized. This is important because human interpretation of CXR is only moderately specific for PTB, has variable sensitivity, is marked by limited inter-reader reliability, and the reproducibility is limited (7, 28). Therefore, the use of a radiologist as the reference standard represents a potential source of information bias.

Lastly, an evaluation of a diagnostic test should also encompass a consideration of the social context in which the test is meant to be used. This includes ethical considerations, legal issues, and economic costs and benefits (22). This evaluation remains out of the scope of this thesis, but highlights a critical area that requires further research. The WHO's strategic plan for the use

of CXRs in TB care highlights the importance of determining existing infrastructure, availability of human resources, and assessing if local stakeholders are interested in using CXRs before recommending widespread uptake and use (12).

Meta-analysis

A quantitative systematic review, or meta-analysis, is often incorporated within systematic review whereby statistical methods are used to combine the results of two or more studies (21). Meta-analysis results are generally more precise, but conducting a meta-analysis is not always appropriate. According to Cochrane, it is inappropriate to conduct a meta-analysis under the following conditions: (a) there is a large degree of missing information; (b) there is unexplained heterogeneity between studies that make an average effect difficult to interpret; (c) there are large differences in populations, interventions, comparisons, or methods used in the studies making an average effect across studies meaningless (29).

Appropriately, a meta-analysis was not conducted in the previous review of CAD4TB (23). One overarching concerning was that several CAD4TB versions have been studied-- 1.08, 3.07, 4.1, 5, but because abnormality scores were not equivalent across the versions, the evidence-base for each was quite small. Another major limitation of CAD4TB is that it is unclear what threshold score should be used to optimize diagnostic accuracy. In the majority of studies, investigators either (i) reported the diagnostic accuracy across a range of threshold scores, or (ii) selected a threshold score post-hoc to match the accuracy of the study's human readers, or (iii) chose a training set of CXRs from the larger validation set to determine which threshold score to use. Other methodologic concerns included the potential for selection bias due to the inappropriate exclusion of participants; limited generalizability due to the evaluation of CAD4TB using CXR from a dataset that also contributed to training the software; and the involvement of the software's developers in the publications. We aimed to conduct a meta-analysis in our review only if the above the methodological concerns could be addressed.

AIM 2

Research has shown that study design features will affect the reported diagnostic test accuracy in studies evaluating new diagnostic techniques (30-32). Methodology features known to affect diagnostic accuracy include, but are not limited to the following: selection bias, whether data was collected retrospectively or prospectively, sampling procedures, non-blinded interpretation of the index test, and post analysis definition of a cut-off scores (31). In an effort to improve the quality of reporting of new diagnostic tests, guidelines have been developed. The Standards for Reporting of Diagnostic Accuracy (STARD) statement provides a reporting checklist to allow clinicians and researchers to more easily assess the quality of the studies and their results and improve methodology (33). However, despite these guidelines, the quality of reporting has remained similar to the pre-STARD standards era, which dates back to 2003 (32).

Therefore, we sought to asses and quantify how methodologic factors specially affected reported CAD accuracy results based on the degree of patient selection, index test used, and reference standard bias. Over optimism of new diagnostic tools has also been shown to effect accuracy measures (34). Therefore, we also compared clinical studies to development studies reported accuracies. Clinical studies primarily focused on the assessment of the accuracy of an already-developed CAD software. Development studies, primarily focused on the early methods for creating a CAD and an assessment of diagnostic accuracy.

Additionally, CAD for TB detection has unique study features which may impact its accuracy. These include the type of AI software used, how the AI software was trained and tested, and the refence standard employed. As described above, AI procedures are evolving, and now include ML and DL techniques. While similar, they have important computational differences and therefore may have different accuracy results. Furthermore, it is important to know how these AI tools were trained and tested. ML and DL require data sets to train their prediction models, which then need to be tested or validated. Ideally, different data should be used for training and testing, IE. different CXRs, as the ultimate goal is to generalize the predictive

capacity. Estimating the accuracy (sensitivity, specificity, AUC) on data which the model was fitted (trained) will lead to an overestimation of the predictive power.

Lastly, the reference standard use for TB diagnosis will also impact the reported accuracy. As previously described, CXR alone should not be used for PTB diagnosis. Microbiologic testing of sputum remains the gold standard for PTB identification. Studies that used human interpretation of CXR as the reference standard are susceptible to information bias. Therefore, we compared accuracies of studies that used microbiology vs human interpretation of CXR reference standards.

Chapter 3: Figures and Tables

Statistical measures of diagnostic test utility		
Statistical measure	**Definition**	**Calculation**
Accuracy	The proportion of people correctly identified as either having or not having the disease	(TP + TN) / (TP + FP + FN + TN)
Sensitivity	The proportion of people who have the disease who test positive	TP / (TP + FN)
Specificity	The proportion of people who do not have the disease who test negative	TN / (FP + TN)
Positive predictive value	The proportion of people who test positive and who have the disease	TP / (TP + FP)
Negative predictive value	The proportion of people who test negative and who do not have the disease	TN / (FN + TN)
Positive likelihood ratio	How likely a positive test result is in people who have the disease as compared to how likely it is in those who do not have the disease	Sensitivity/(1-specificity)
Negative likelihood ratio	How likely a negative test result is in people who have the disease as compared to how likely it is in those who do not have the disease	(1-sensitivity)/specificity

Table 1. Definition and calculation of statistical measures used to express diagnostic test utility

TP, true positive; TN, true negative; FP, false positive; FN, false negative

Connecting text

To answer our PICO question from Aim 1, we completed a systematic review, however, the above noted methodological concerns identified in the previous systemic review still applied and therefore a meta-analysis was not pursued. We synthesized the data according to the Cochrane Effective Practice Organization of Care (EPOC) recommendations for “synthesizing results when it does not make sense to do a meta-analysis". We reported standardized outcome measures of sensitivity and specificity, calculated the interquartile ranges (IQR) of these measures, and included plain language summaries of the accuracy results. We categorized the accuracy measures by the type of CAD study (clinical vs. development) and by screening or triage use-case.

CHAPTER 4: *A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis*

Miriam Harris, Amy Qi, Nazi Torabi, Dick Menzies, Alexei Korobitsyn, Madhukar Pai, Ruvandhi R. Nathavitharana, Faiz Ahmad Khan

Introduction

The need to improve tuberculosis (TB) diagnostic and screening services in high-burden countries is clear. In 2016, active TB was the leading cause of death due to an infectious agent, and only 69% of the 10.4 million people that developed this disease were detected by or notified to national TB programmes (10). In the absence of a rapid, cheap, accurate and ideally non-sputum based TB diagnostic test, the World Health Organization (WHO) and other stakeholders have identified the need for a triage test to determine which people should undergo confirmatory TB diagnostic testing. In developed countries, chest x-rays (CXRs) have been used for the evaluation of persons presenting with symptoms of possible active TB (triage), and for screening of asymptomatic individuals in high risk groups, for several decades (1). However, their uptake in high TB burden countries, particularly in resource-constrained settings, has been limited (2, 3).

In recent years, there has been increasing interest in expanding access to chest radiography in order to improve TB case detection in high-burden areas (7). However, one of the challenges is the paucity of professionals to interpret radiographic images in resource-constrained settings (4). Computer technology has developed the ability to automate intellectual tasks, also known as artificial intelligence (AI). Recently, AI has been applied to the analysis of CXR images to identify abnormalities—referred to as computer-aided detection, or CAD— and represents one potential solution the personnel shortage. CAD uses two types of AI for CXR reading: Machine learning (ML) and Deep Learning (DL). ML is a type of AI analysis that relies less on human

specification (i.e. defining a set of variables to be included) and instead allows algorithms to decide what variables are important (8, 9). DL is a subset of ML which attempts to model brain architecture (9). It uses neural networks, or overlaying models, that emphasize learning increasingly meaningful representations of the data (9). The World Health Organization (WHO) has called for greater evidence before endorsing the use of CAD in TB diagnostic and screening pathways (7).

To date, only one systematic review has been undertaken of CAD for TB (35), and was limited to reviewing the only commercially available software at the time of publication. Amongst the 5 included studies, the reviewers identified a number of methodological limitations that could have resulted in potential bias in estimating diagnostic accuracy, limited the generalizability, and this prevented pooling of results. Because the prior review was limited to studies of the single commercially available software, it excluded the vast majority of studies of CAD for detecting PTB. Hence, in order to provide a more comprehensive and expansive summary of the CAD literature we undertook an updated systematic review which included non-commercially available CAD studies. Our primary objectives were to evaluate the evidence base with regards to the estimation of the diagnostic accuracy of CAD, including assessing potential for bias, and if appropriate, to calculate pooled estimates of area under the receiver operating characteristic curves (AUC), sensitivity, and specificity. Secondary objectives were to evaluate study-level factors associated with diagnostic accuracy; including those related to the design of the study, and the type of software used (ML versus DL).

Methods

Design

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (Appendix) (36). The International Prospective Register of Systematic Reviews (PROSPERO) registration number of this protocol is CRD42018073016.

Date source and search strategy

A search strategy was developed in consultation with an academic librarian (NT) to identify published articles in MEDLINE (Ovid), EMBASE (Ovid), PubMed, and Scopus (Appendix). The search strategies included subject headings (where applicable) and text words for the concepts of pulmonary TB (PTB), computer aided diagnosis, and diagnostic accuracy. As CAD is a modern technology the search period was limited to papers published after January 1, 2005, and included articles published up to November 28th, 2017. Studies were limited to English and French.

Study Selection

We included all published studies that used any form of computer software to analyze CXR in place of human readers, for PTB detection purposes. Studies were excluded if they reported CAD for diagnostic imaging other than CXR, or if CAD was used for diseases other than PTB. Studies reported only in conference abstracts were excluded. Two independent reviewers selected studies for inclusion (MH, AQ). Conflicts were reviewed by a third reviewer (FAK).

Data extraction

Data were extracted using a standardized extraction form (Appendix). Two reviewers performed the extraction, with one reviewer (MH) verifying all data forms completed by the second reviewer (AQ). Data collected included year of enrollment; funding sources and conflicts of interest; software name and version number; country where study was completed; CXR site and number on which the software was trained; model of CXR machine, and digitization methods; study design and patient selection methods; inclusion and exclusion criteria; microbiologic tests collected; scoring of software tools and methods of scoring selection; patient characteristics including HIV status, age, and history of TB; and diagnostic accuracy measures including sensitivity, specificity, AUC for microbiologic and radiologic references.

Descriptive analysis

We classified studies as either Development or Clinical. Development studies primarily focused on reporting methods for creating a CAD program for PTB, and some included an assessment of diagnostic accuracy—the latter being the focus of our systematic review. Development studies were often published in engineering, computer science, medical imaging journals, or proceedings from engineering or medical imaging conferences. The development studies were further subdivided based on the type of AI technology used (ML versus DL).

Clinical studies primarily focused on the assessment of the accuracy of an already-developed CAD software. We further classified clinical studies based on the context in which the CXR was used, using WHO terminology for categorizing usage of x-ray as either for Triage or for Screening.(7) In Triage studies, CXRs were used in a healthcare setting—hospital, or clinic—as part of the diagnostic pathway of someone with PTB symptoms. In Screening studies, CXRs were used for active case finding or prevalence surveys, where populations are screened to identify those with active TB often regardless of symptoms. The distinction was made because the prevalence of more advanced or extensive disease will be higher in the Triage setting, thereby affecting the sensitivity of CXR and hence the accuracy of CAD.

Quality assessment with respect to the evaluation of diagnostic accuracy

The data sources used for evaluating diagnostic accuracy of CAD were databases consisting of CXRs, with each image linked to a reference standard result classifying PTB as present or absent. Some of these data sources had been used by more than one Development study. We evaluated these data sources for potential risk of bias by applying a modified Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 approach (18). As our interest was to assess the composition of the database itself including how PTB cases were defined, we restricted our approach to the domains of patient selection and the reference test. Because development studies often did not provide sampling or reference details about the data sources, we sought additional information from citations that described the data sources (37-40).

We applied QUADAS-2 to all the CAD studies, assessing each study across the four domains (patient selection, the performance of the index test, performance of the reference test, and flow and timing). In all quality assessments, when the reference standard used for determining a CAD program's diagnostic accuracy was image interpretation by a human reader instead of microbiologic testing of sputum, we judged this as a potential source of bias. This is because human interpretation of CXR is moderately specific for PTB, has variable sensitivity, is marked by limited inter-reader reliability, and the reproducibility is limited (7, 28).

Statistical analysis

Diagnostic accuracy measures (sensitivity, specificity, AUC) were reported when available. For the studies that reported sensitivities and specificities, if two by two tables were not available, we back calculated counts based on reported accuracy measures to build forest plots. A meta-analysis was not undertaken given that different software programs were used, and for most studies the raw data necessary to meta-analyze diagnostic accuracy measures were unavailable. For studies of the most commonly reported software, CAD4TB, a meta-analysis was also not pursued due to the variability of the methods and versions tested.

The following study-level factors were evaluated as potential determinants of the reported AUC: type of CAD study (Development vs Clinical); the method of AI software (ML versus DL); whether the same CXRs used for evaluating diagnostic accuracy were the same CXRs that had been used to train the software; the type of reference standard for PTB (microbiologically confirmed vs human interpretation of CXR image); and the degree of patient selection, index test, and reference standard bias. While the data were insufficient for a traditional meta-analysis, to identify associations between these factors and reported AUC, we compared the pooled distribution of the reported AUCs between groups defined by these study-level factors using Kruskal-Wallis tests. This assessment was done for the AUC but not for Sensitivity or Specificity, as the latter two were reported in too few studies to undertake a meaningful comparison of distributions.

For all clinical studies and development studies which reported sensitivity, specificity, and true positives, forest plots were used to visually assess heterogeneity of diagnostic accuracies.

Results

Study selection

We identified 2697 unique citations (Figure 4), of which 2349 studies were excluded at the title and abstract phase. Of the remaining 348, 306 were excluded after full-text review. Amongst the 42 included articles, 32 were classified as development studies and 10 were classified as Clinical (Table 2). The software developers were either authors or funded the research in 7/10 (70%) of the clinical studies (20, 41-46), and in 100% (32/32) of the Development studies.

Overview of studies

Within the Development studies, 4/32 (13%) employed DL methods while the remaining 28/32 (87%) used ML approaches (Table 2) (15, 16, 47-76). An important consideration when evaluating the accuracy of a CAD software, is that it should be tested using a set of CXR images that are separate from the training set (i.e. avoid testing accuracy with CXRs that were used for training, or CXRs that were not used for training but that originate from the same subset/study as those with which the program was trained). Otherwise, the evaluation is likely to overestimate the diagnostic accuracy, and will also have limited generalizability (17). Within the development studies that reported accuracy measures, 3/25 (12%) did not report the database used to train and test their software. The majority, 22/25 (88%), used the same databases to train and test their software (Table 3). For the majority of development studies demographic data were not reported or
found.

All clinical studies used the CAD4TB software, which is a ML-based program. Within the triage use-case studies, 5/7 (71%) used a microbiologic reference standard on all participants (20, 42, 45, 77, 78). Within the screening studies, 2/3 (67%) used a microbiologic reference (43, 79). In

two clinical studies CAD4TB itself was used to select which participants underwent microbiologic testing, hence the software's diagnostic accuracy could not be assessed (41, 79). Demographically speaking, the study populations of all the triage studies with microbiologic references were quite similar (Table 4). Notably, the estimated HIV and TB prevalence in the Triage studies were quite high, ranging from 33% to 68%, and 18% and 33% respectively. The screening studies had lower TB prevalence, as expected (Table 4).

Quality assessment development studies

We first assessed the databases that were used as sources of CXR images and reference standards in development studies (Table 5). Risk of selection bias was high in 2/16 (12%) of the databases. One dataset did not include PTB cases, and the other only included patients with "typical TB" images (38, 70). Selection bias was unclear in 6/16 (38%), and low in 8/16 (50%) where consecutive enrollment either prospectively or retrospectively was used. The reference standard risk of bias was high in 9/16 (56%) studies as a human reader was used, unclear in 3/16 (19%), and low in 4/16 (25%) where a microbiologic reference was used.

The quality of the development Studies with respect to the assessment of diagnostic accuracy is reported in Table 6. Selection biased was largely determined by which databases were used (Table 5). There was a high degree of selection bias in 7/25 (28%) studies (54, 58, 68, 70, 80, 81), it was unclear in 13/25 (52%) studies, and was low in 6/25 (24%) studies. There was a high risk of bias in 100% of the studies in the assessment of the index test, as no CAD programs had pre-specified threshold scores and these were set after the analysis. Additionally, 23/25 (92%) of the studies were considered to have a high degree of bias and low degree of applicability with regards to the reference test utilized due to use of a human reader's interpretation of CXRs. The flow and timing had low bias in 13/25 (52%) studies, in 11/25 (44%) it was unclear, and in 1/25 (4%) it was high.

Quality assessment of clinical studies

Table 7 summarizes the QUADAS-2 assessment of the clinical studies. All triage studies used a consecutive enrollment strategy, with 3/7 (43%) being prospective, 4/7 (57%) retrospective (Table 8). There were methodological concerns that likely resulted in a high degree of selection bias in 4/10 (40%) of the studies (20, 44, 46, 78). This was secondary to case-control design (44), and inappropriate exclusion of patients in the analysis (20, 46, 78). CAD4TB cut off score was pre-specified in only 4/10 (40%) of the studies (41, 42, 45, 82). The remainder of the studies developed threshold scores post analysis using the receiver operator curves (ROC), therefore were determined to have a high risk of bias (20, 43, 44, 46, 78, 83). The majority of studies, 7/10 (70%) had low bias with regards to the use and performance of the reference standard (20, 42, 43, 45, 46, 83). In two studies, the reference standard was pre-selected to be done by CAD4TB, and therefore were determined to have a high degree of bias (41, 82). In another study, the reference standard was human reading of the CXR which was deemed to have a high risk of bias (44). The flow and timing had a high risk of bias in 2/10 (20%) of the studies due to CAD4TB selection of the reference standard,(41, 82) was unclear in 3/10 (30%), and low in 5/10 (50%).

Diagnostic accuracy reported in Development Studies

We found 25/32 (78%) of the development studies reported measures of accuracy for index tests. Of the 25 references that did include accuracy assessments, the AUC ranged from 0.86 to 0.99, sensitivity from 0.56 to 0.97, and specificity from 0.36 to 0.95 (Table 3). The forest plots graphically display the diagnostic heterogeneity of the sensitivity and specificity of the development studies that published sensitivity, specificity, and the number of true positive TB cases (Figure 5).

Diagnostic accuracy reported in clinical Studies

The forest plots graphically display the diagnostic heterogeneity of the sensitivity and specificity of the triage CAD4TB studies that used a microbiologic reference (Figure 5). In the triage studies, the sensitivity ranged from 0.85 to 1.00, and specificity ranged from 0.23 to 0.69. There

was only one screening study that used microbiologic reference and it had a sensitivity 0.53 and specificity of 0.98 (43). The other screening study (44) compared CAD4TB to a radiologic reference interpreted by a human reader and demonstrated a sensitivity of 0.59 and specificity of 0.78. Notably, the sensitivity of CAD4TB was higher when using NAAT as the microbiologic reference standard compared to culture. Given the methodological heterogeneity, the lack of a standardized cut off scores, and the variability of software versions used, a meta-analysis was not completed.

Assessment of study-level factors associated with reported AUC

Figure 6 shows the distribution of reported AUCs with studies stratified by study level characteristics. Reported AUCs were higher in: development studies (median [IQR] AUC: 0.87 [0.81-0.90]) versus clinical studies (0.68 [0.65-0.75]; p-value 0.004); and with programs created using DL (0.93 [0.90-0.97]) versus ML (0.81 [0.69-0.88]; *p*=0.001). While not statistically significant, we found that the mean AUC of studies using human readers as the reference standard were higher than those studies using a microbiologic reference standard of 0.87 [0.81-0.90] versus 0.75 [0.65-0.83] respectively (p=value of 0.067). There was no significant difference in AUCs of studies that used the same CXRs as the source for software development and evaluation of diagnostic accuracy, or of the AUCs by the degree of patient selection, index test, or reference standard bias (Figure 6).

Discussion

In this systematic review, we sought to determine the diagnostic accuracy of CAD software programs for detecting PTB on CXRs, but due to study heterogeneity, we were unable to meta-analyze the data. We identified a number of methodological limitations in the existing evidence base that limits the use of this technology in clinical settings. Moreover, we identified a number of study-level factors associated with the reported accuracy, which should be taken into consideration when evaluating future CAD studies.

The majority of the CAD evidence base for PTB detection consists of development studies. While many reported some measure of diagnostic accuracy, this was done without reporting the potential risk of bias. Applying a widely accepted standardized tool—QUADAS-2— for evaluating the quality of diagnostic studies we found that the potential risk of bias was common in these databases, or could not be assessed. We suggest future development studies apply the QUADAS-2 tool to assess for bias of the databases (Box 1).

The clinical studies evaluated the only currently commercially available software, CAD4TB. As noted above, meta-analysis was not completed due to the methodological heterogeneity, the lack of standardized cut off scores, and the variability of software versions used. While the software achieved high sensitivities (0.85-1.0), there was a large degree of variability in the reported specificities (0.23-0.69). Furthermore, the analysis in some studies was performed on CXRs from datasets or sites that may have also contributed to training the software, likely resulting in an overestimation of the predictive power. Lastly, because the populations studied had very high HIV and TB prevalence, the results have limited generalizability to other populations.

We identified a number of study-level factors that were associated with the reported AUC. These included the type of technology used to classify images, and whether it was a Development or Clinical study. The accuracy of DL vs ML studies was higher (mean AUC DL vs ML p-value 0.001), suggesting superior diagnostic accuracy of DL technology. The mean AUC of development studies was higher than clinical studies (p-value 0.002). This likely because of the greater risk of bias due to the lack of pre-specified cut off scores, the use of the same databases for training and testing, and the use of a human reader as the reference standard. We found a signal that studies which used a human reference standard compared to a microbiologic had a higher mean AUC (p-value 0.067). Human interpretation of CXR should not be used as the sole diagnostic test for PTB due to its poor accuracy, and therefore studies which used a human reader reference standard likely systematically overestimated the diagnostic utility of their software. We did not find a significant difference in AUCs from studies that used the same CXRs

for training and testing. However, we can extrapolate from other studies that using the same databases for training and testing will results in the systematic overestimations of reported predative value (84).

We suggest that future studies ensure that software being tested in a valid fashion that will allow for clinical applicability. This includes a description of how CXRs were selected for training and testing. Furthermore, we suggest that different CXRs from separate databases be used for training and testing. Studies should also clearly describe how a true positive TB case is defined, and ideally this should be done using a microbiologic reference standard. Lastly, if the software has a continuous output, a cut-off or threshold score to define a positive or negative case should be reported along with how this was determined (Box 1). The US Food and Drug Administration (FDA) requires all of these standards be met and additionally necessitates clear instructions for clinical use in order to become approved according to their guidelines of CAD applied to radiology devices (17).

One potential weakness of this review is that we only included studies from the published literature, which could increase the risk that publication bias affected our reported results. Additionally, we restricted our search to English and French studies only. Furthermore, we were unable to complete a meta-analysis of the clinical studies due to the potential bias and heterogeneity in the references. Therefore, we are unable to comment on the pooled accuracy of CAD.

This systematic review highlights the need for additional research of CAD CXR reading for PTB identification. To strengthen the evidence-base using existing data, an individual patient data (IPD) meta-analysis, conducted independently of CAD4TB's developers, could be used to assess the diagnostic accuracy of CAD4TB. To our knowledge, this is the first study to analyze the quality of current CXR databases that have been used to train and test multiple CAD software tools. While CAD4TB achieved high sensitivities, with variable specificities, methodological limitations within the existing literature limits our ability to comment on the clinical utility. We

conclude that AI based CAD programs are promising, but more clinical studies are needed that minimize sources of potential bias to ensure validity of the findings outside of the study setting.

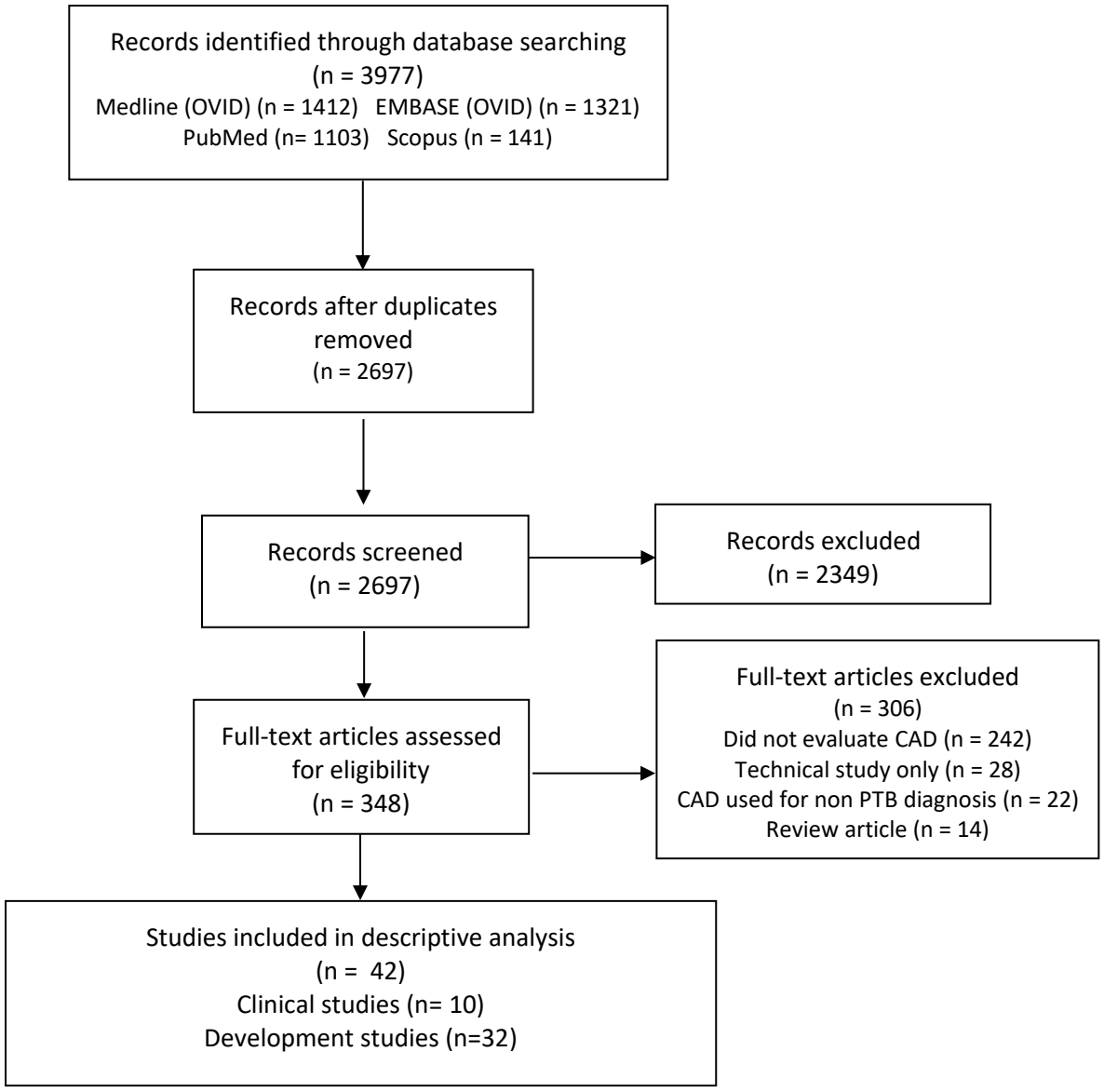

Figure 4. Study flow diagram
Computer aided detection (CAD)

Author and year	Country where CXR completed	Databases used	Computer software	Reference standard	Accuracy measures
Development Studies					
Deep learning					
Lakhani et al, 2017	USA, China	MC, Shenzhen, TJH, Belarus	AlexNet and GoogLeNet	Human reader	AUC, Sn, Sp
Lopes et al, 2017	USA, China	MC, Shenzhen	Not named	Human reader	AUC
Santosh et al, 2016	USA, China	MC, Shenzhen	Not named	Human reader	AUC
Hwang et al, 2016	South Korea, USA, China	KIT, MC, Shenzhen	Alexnet	Human reader	AUC
Machine learning					
Fatima et al, 2017	USA	MC	Not named	Human reader	Sn, Sp
Udayakumar et al. 2017	USA, China	MC, Shenzhen	SVM and CBC techniques	Human reader	AUC
Ding et al, 2017	China, India, Kenya	Kenya dataset, New Delhi dataset, Shenzhen	Not named	Human reader	NR
Hogeweg, et al, 2017	Japan, Sub-Saharan Africa	JSRT, Sub-Saharan Africa	Not named	Human reader	AUC
Maduskar et al, 2016	Zambia	Large Zambian dataset	Not named	Human reader	AUC
Poornimadevi et al, 2016	Japan, USA	JSRT, MC	Not named	Human reader	Sn, Sp
Karargyris et al, 2016	China, Japan	JSRT, Shenzhen	Not named	Human reader	AUC
Melendez et al, 2016	Zambia	Zambian dataset	Not named	Human reader	AUC
Melendez et al, 2015	Zambia, Tanzania, Gambia	Zambian dataset, Tanzania dataset, Gambian dataset	Not named	Human reader	NR
Hogeweg et al, 2015	UK, South Africa	F&T, TB-NEAT	Not named	Human reader, Liquid culture, composite reference standard **	AUC, Sn, Sp
Giacomini et al, 2015	Brazil	Prospective, study-specific[†]	Not named	Liquid culture[+]	NR
Jaeger et al, 2015	China	Shenzhen	Not named	Human reader	NR
Requena-Mendez et al, 2015	Peru	CXR from DOT study in Peru	Not named	Human reader	NR
Jaeger et al, 2014	China, USA, Japan	JSRT, MC, Shenzhen	Not named	Human reader	AUC, Sn, Sp
Melendez et al, 2014	Zambia, South Africa	Zambian dataset	TB-Xpredict	Human reader	AUC
Chauhan et al, 2014	India	New Delhi dataset	Not named	Human reader	NR
Seixas et al, 2013	Brazil	Clinical data set from another study*	Artificial Neural Network	Composite reference**	NR
Sundaram et al, 2013	Not specified	Not specified	Not named	Human reader	NR
Jaeger et al, 2012	USA, Japan	JSRT, MC	Not named	Human reader	AUC
Xu et al, 2011	Japan, Canada	JSRT, Calgary dataset	Andrews' curve	Human reader	TP, FP, FPR
Noor et al, 2011	Malaysia	Retrospective non-clinical study specific radiological dataset	Not named	Human reader	Sn, Sp
Shen et al, 2010	Canada	JSRT, Calgary dataset	Not named	Human reader	TP, FPR

Mouton et al, 2010	South Africa	Clinical dataset from previous study not specific to PTB	Not named	Human reader	AUC
Hogeweg et al, 2010	Sub-Saharan Africa	Sub-Saharan Africa	CAD with rib suppression	Human reader	AUC
Hogeweg et al, 2010	Not specified	Not specified	Not named	Human reader	NR
Lieberman et al, 2009	China	Prospective, study-specific†	Not named	Human reader	NR
Arzhaeva et al, 2009	Netherlands	F&T	Not named	Human reader	AUC
Noor et al, 2005	China, USA	MC, Shenzhen	Andrews' curve	Composite reference**	NR
Clinical studies					
Machine learning					
Rahman et al, 2017	Bangladesh	Prospective, study-specific†	CAD4TB (v 3.07)	NAAT	AUC, Sn, Sp
Melendez et al, 2017	Zambia	Zambia National TB Prevalence Survey	CAD4TB (v 5.0)	Human reader CXR-, Liquid culture/NAAT for CXR+	AUC, Sn, Sp
Muyoyeta et al, 2017	Zambia	Prospective, study-specific†	CAD4TB (v 1.08)	NAAT for CXR+, AFB Smear for CXR-	NR
Melendez et al, 2016	South Africa	TB-NEAT collaborative study	CAD4TB (v 3.07)	Liquid culture	AUC, Sn, Sp
Philipsen et al, 2015	South Africa	TB-NEAT collaborative study	CAD4TB (v 3.07)	NAAT, liquid culture	AUC, Sn, Sp
Steiner et al, 2015	Tanzania	TB REACH project	CAD4TB (v 3.07)	Human reader	AUC, Sn, Sp
Muyoyeta et al, 2015	Zambia	Prospective, study-specific†	CAD4TB (v 1.08)	NAAT, AFB Smear for CXR-	AUC, Sn, Sp
Breuninger et al, 2014	Tanzania	TB Cohort and TB CHILD study	CAD4TB (v 3.07)	Liquid culture, AFB smear	AUC, Sn, Sp
Muyoyeta et al, 2014	Zambia	Prospective, study-specific†	CAD4TB (v 1.08)	NAAT	AUC, Sn, Sp
Maduskar et al, 2013	Zambia	Prospective, study-specific†	CAD4TB (v 1.08)	Liquid culture, AFB smear	AUC, Sn, Sp

Table 2. Methods of studies included in the descriptive analysis

CXR, chest x-ray; USA, United States of America; UK, United Kingdom; AI, artificial intelligence; MC, Montgomery County; TJH, Thomas Jefferson Hospital dataset; JSRT, Japanese Society of Radiology; KIT, Korean Institute of Tuberculosis; F&T, Find and Treat; SVM, *Support vector machines;* CBC, clustering based classification; CAD, computer aided detection; NAAT, nucleic acid amplification test; AFB, acid fast bacilli; '+', positive; '-', negative; AUC, area under the receiver operating curve; Sn, sensitivity; Sp, specificity; NR, not reported; TP, true positives; FP, false positives; FPR, false positive rate

* Trajman et al. Pleural fuid ADA, IgA-ELISA and NAAT sensitivities for the diagnosis of pleural tuberculosis Study

**Composite reference: positive culture/NAAT and/or initiation of TB treatment

†In these studies the study database was developed prospectively for the specific study

Author and year	Database(s) used for training of CAD	Number of CXRs used for training	Database(s) used for testing CAD	Number of CXRs used for testing	Number of TB positive CXR	AUC (95% CI)	Thres-hold score	Sn (95% CI)	Sp (95% CI)
Deep learning									
Lakhani et al, 2017	MC, Shenzhen, TJH, Belarus	857	MC, Shenzhen, TJ, Belarus	150	75	0.99 (0.96-1.00)	NR	0.97 (0.90 -1.0)	0.95 (0.87-0.98)
Lopes et al, 2017	NR	NR	Shenzhen MC,CI,NR	1031	550	0.834 (Shenzhen) 0.926 (MC)*	NR	NR	NR
Santosh et al, 2016	NR	NR	Shenzhen MC,CI,NR	878	400	0.93 (Shenzhen) & 0.88 (MC)*	NR	NR	NR
Hwang et al, 2016	KIT	9221	KIT, MC, Shenzhen	2427	NR	0.96*+	NR	NR	NR
Machine learning									
Fatima et al, 2017	MC	138	MC	138	58	NR	NR	0.83*	0.78*
Udayakumar et al.	MC, Shenzhen	NR	MC, Shenzhen	NR	NR	0.87*	NR	0.81*	0.74*
Hogeweg, et al, 2017	JSRT, Sub-Saharan Africa	NR	Sub-Saharan Africa	348	174	0.891*	NR	NR	NR
Ding et al, 2017	NR	NR	Kenya dataset, New Delhi dataset, Shenzhen	NR	NR	0.949 (China), 0.982 (India), 0.76 (Kenya)*	NR	NR	NR
Maduskar et al, 2016	Large Zambian dataset	629	Large Zambian dataset	638	NR	0.9*	NR	0.83*	0.70*
Poornimadevi et al, 2016	JSRT	247	JSRT	247	NA	NR	NR	0.56*	0.36*
Karargyris et al, 2016	Shenzhen	43	JSRT, Shenzhen	NR	NR	0.93*	NR	NR	NR
Melendez et al, 2016	Zambian dataset	461	Zambian dataset	456	248	0.87*	0.45	NR	NR
Melendez et al, 2015	Zambian dataset, Tanzania dataset, Gambian dataset	Zambia (461) Tanzania (435) Gambia (427)	Zambian dataset, Tanzania dataset, Gambian dataset	Zambia (456) Tanzania (434) Gambia (423)	Zambia (248) Tanzania (226) Gambia (197)	0.86 Zambia, 0.88 Tanzania, 0.91 Gambia*	NR	NR	NR
Hogeweg et al, 2015	F&T, TB-Neat	F&T (200), TB-Neat (200)	F&T, TB-Neat	F&T (200), TB-Neat (200)	F&T (87), TB-Neat (66)	F&T micro 0.87 (0.81-0.92), RD 0.85 (0.79-0.91), TB-Neat RD 0.90 (0.85-0.94) (0.69-	NR	NR	NR

						0.83) micro 0.74*##			
Jaeger et al, 2014	MC, Shenzhen, JSRT	1000	MC, Shenzhen	753	333	0.87*	NR	0.78 (0.70 - 0.85)	0.81 (0.71-0.89)
Melendez et al, 2014	Zambian dataset	461	Zambian dataset	456	NR	0.88*	NR	NR	NR
Chauhan et al, 2014	New Delhi dataset	204	New Delhi dataset	102	153	Gist 0.96 (0.86-0.99) DA, Gist 0.89 (0.77-0.96) DB##	NR	0.96 DA, 0.88 DB*	0. 92 DA, 0.84 DB*&
Sundaram et al, 2013	NR	95	NR	95	52	NR	NR	0.75*	0.90*
Jaeger et al, 2012	JSRT	247	MC	138	NR	0.83*	NR	NR	NR
Xu et al, 2011	JSRT, Calgary dataset	60	JSRT, Calgary dataset	60	NR	NR	NR	0.68*	0.68*
Noor et al, 2011	Retrospective non-clinical study specific radiological dataset	90	Retrospective non-clinical study specific radiological dataset	213	208	NR	NR	0.88*	0.84*
Shen et al, 2010	JSRT, Calgary dataset	18	JSRT, Calgary dataset	131	19	NR	NR	0.82*	NR
Mouton et al, 2010	Clinical dataset from previous study not specific to PTB	119	Clinical dataset from previous study not specific to PTB	119	NR	NR	0.78*	NR	NR
Hogeweg, et al, 2017	CRASS database	348	CRASS database, JSRT	498	NR	0.75*	NR	NR	NR
Arzhaeva et al, 2009	F&T	217	F&T	217*++	37	NR	0.83 TB suspect TB, 0.74 micro TB*###	NR	NR

Table 3. Accuracy measures reported by development studies

CAD, Computer aided detection; MC, Montgomery County; TJH, Thomas Jefferson Hospital dataset; NR, not reported; KIT, Korean Institute of Tuberculosis; JSRT, Japanese Society of Radiology; F&T, Find and Treat; AUC, area under the receiver operating curve; 95% CI, 95 percent confidence interval; NR, not reported; DA, dataset A; DB, dataset B; Sn, sensitivity; Sp, specificity;; TP, true positives; FP, false positives; FPR, false positive rate;

* No 95% CI reported +Average AUC from KIT, MC, Shenzhen

++ 128 of the normal images were the same CXRS used in the training

An external and radiological reference standard were used. The external reference for tuberculosis was set by an independent test not associated with the CXR; the result of a sputum culture testing for the TB-NEAT database and a combination of sputum culture testing and clinical diagnosis for the Find & Treat database

Two CXR digital image datasets, dataset A and B, were obtained from two different X-ray machines available at the National Institute of Tuberculosis and Respiratory Diseases, New Delh

The database was split between TB suspect cases were re-read by a third radiologist, and if classified differently were excluded. The database contained 256 normal radiographs, 178 TB suspect radiographs, and 37 microbiologically diagnosed TB CXRs.

Reference	All	Gender		Age			Previous TB		HIV		Estimated TB prevalence (%)
		Male (%)	Female (%)	< 15 yrs (%)	≥ 15 yrs (%)	average age	Yes (%)	No (%)	+ve (%)	-ve (%)	
CAD4TB Triage											
Rahman et al, 2017	17066	11368 (67)	5698 (33)	0 (0)	17066 (100)	NR	NR	NR	NR	NR	27.7%
Melendez et al, 2016	392	240 (61)	152 (39)	0 (0)	392 (100)	40	NR	NR	130 (33)	262 (67)	23.3%
Breuninger et al, 2014	861	433 (50)	428 (50)	0 (0)	861 (100)	42	144 (17)	717 (83)	379 (44)	482 (46)	18.6%
Muyoyeta et al, 2014	350	215 (61)	135 (38)	NR	NR	36.5	78 (22)	272 (78)	190 (54)	166 (57)	33.3%
Maduskar et al, 2013	161	119 (74)	42 (26)	0 (0)	161 (100)	35.8	NR	NR	110 (68)	51 (32)	18.3%
CAD4TB Screening											
Melendez et al, 2017	23838	10440 (44)	13398 (56)	0 (0)	23838	36	NR	NR	NR	NR	106 (2.9)
Muyoyeta et al, 2017	919	370 (40)	549 (60)	NR	NR	15	57 (6)	862 (94)	138 (15)	781 (85)	19 (2)

Table 4. Demographics of CAD4TB studies with microbiologic reference standard

CAD, computer aided diagnosis; yrs, years; NR, not reported; TB, tuberculosis; HIV, human immunodeficiency virus

CXR image dataset used for evaluating CAD	Risk of Bias	
	Patient Selection	Reference Test
MC	U	H
Shenzhen	U	H
JSRT*	H	H
Sub-Saharan Africa	L	H
Kenyan dataset	U	U
New Delhi dataset	U	U
Large Zambian dataset	L	H
Zambian dataset	L	H
Gambian dataset	L	H
Tanzania dataset	L	H
F&T	L	L
TB-NEAT	L	L
TJH	U	L
Belarus	U	L
Calgary dataset**	H	H
KIT dataset	U	U

Table 5. Quality assessment of Datasets used to test and train CAD software of Development Studies: risk of bias and applicability concerns

MC, Montgomery County; JSRT, Japanese Society of Radiology; F&T, Find and Treat; TJH, Thomas Jefferson Hospital dataset; KIT, Korean Institute of Tuberculosis; U, unclear; H, high; NA, not applicable; L, low

* JSRT data set does not include PTB cases, but rather comprises images with single pulmonary nodules, confirmed by computed tomography and histology as either benign or pathologic

** Calgary dataset included preselected “typical PTB” images

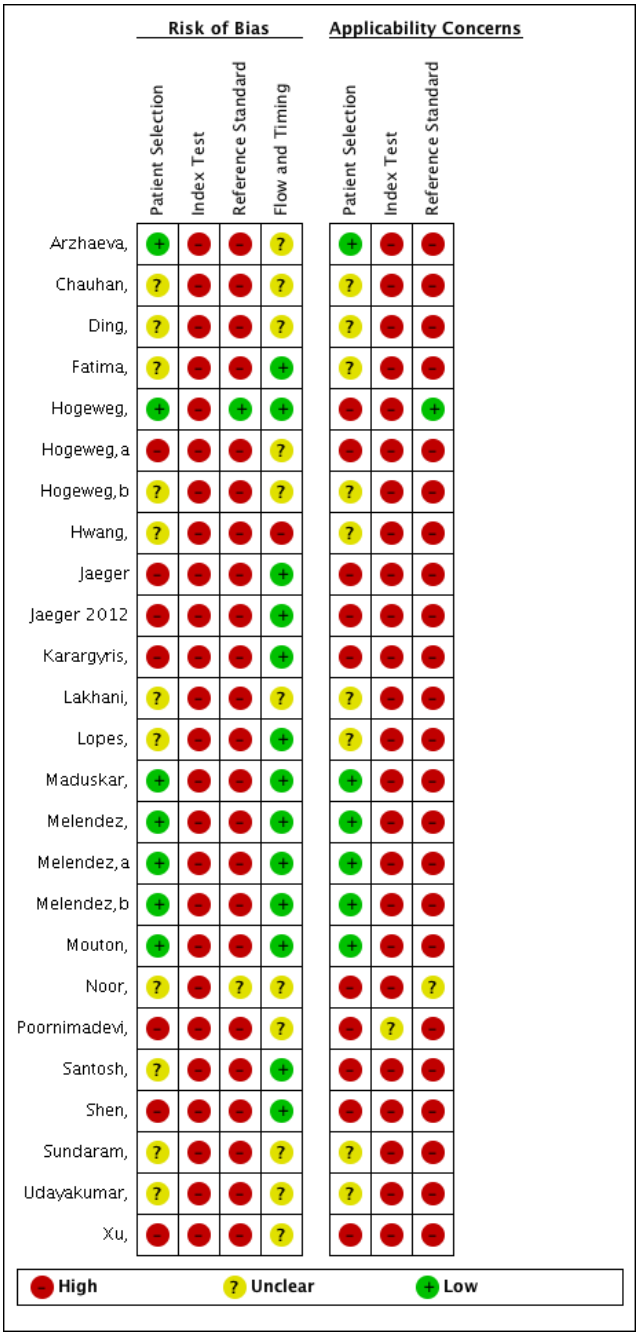

Table 6. Quality assessment (QUADAS 2) graph of development Studies

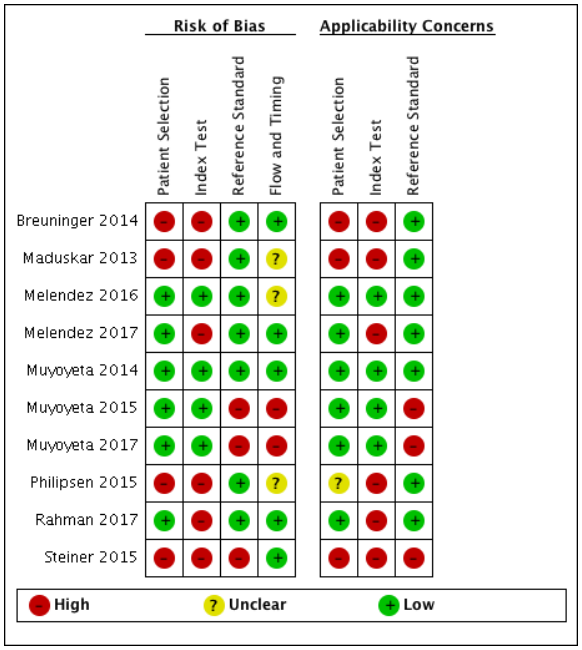

Table 7. Quality assessment (QUADAS 2) graph of clinical Studies

Reference	Eligible, N	Enrolled, N (% of eligible)	Reported for assessment of CAD4TB, N (% of enrolled)	Smear positive, N (% of reported)	Culture or NAAT positive N (% of reported)
CAD4TB TRIAGE					
Rahman et al, 2017	18036	17134 (95%)	17066 (99%)	NR	2623 people (15%)
Melendez et al, 2016	NR	392	392 (100%)	NR	73 (19%)
Philipsen et al, 2015	758	419 (55%)	388 (93%)	NR	133 (34%)
Muyoyeta et al, 2015	10618	9509 (90%)	9482 (99%)	8 (<1%) *	2090 (22%) *
Breuninger et al, 2014	894	861 (96%)	566 (66%)	146 (17%)	194 (23%)
Muyoyeta et al, 2014	458	391 (85%)	350 (90%)	52 (13%)	96 (35%)
Maduskar et al, 2013	NR	161	161 (100%)	69 (43%)	97 (60%)
CAD4TB SCREENING					
Melendez et al, 2017	46099	25805 (56%)	23838 (92%)	NR	106 * *
Muyoyeta et al, 2017	919	865 (94%)	865 (100%)	0 *	19 *
Steiner et al, 2015	516	511 (99%)	511 (100%)	NR	NR

Table 8. Selection, enrolment of CAD4TB studies with microbiologic reference standard

NR, not reported; CAD, computer aided diagnosis; NAAT, nucleic acid amplification test

* Patients with a normal CXR by CAD received an AFB smear, while patients with an abnormal CXR as per CAD received NAAT

** Patients with a abnormal CXR as per radiologist reading, or presumptive TB based on TB symptoms received culture/NAAT

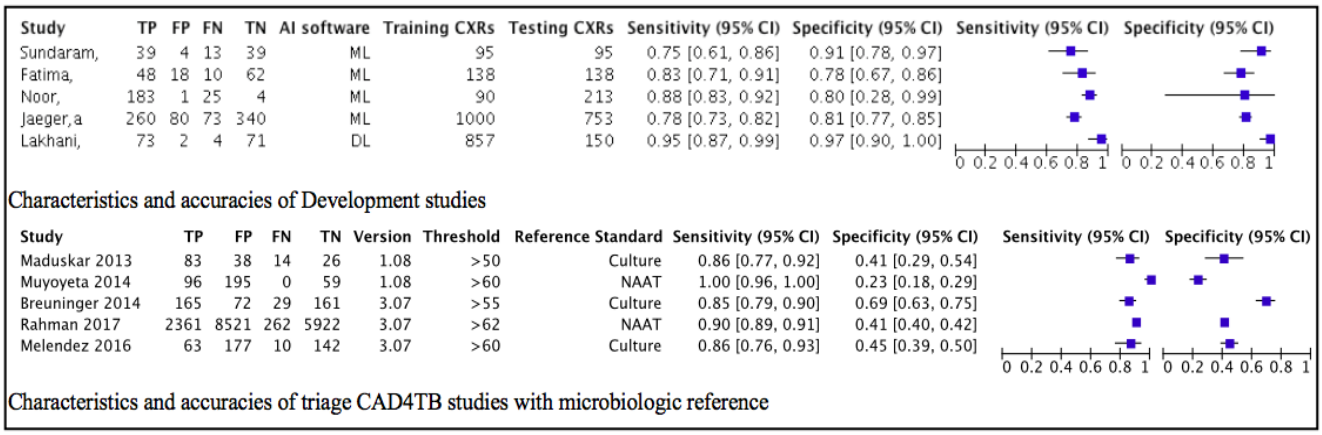

Figure 5. Forest plots of accuracy measures of development and CAD4TB studies

TP, true positive; FP, false positive; FN, false negative; TN, true negative; NAAT, nucleic acid amplification test; CI, confidence interval

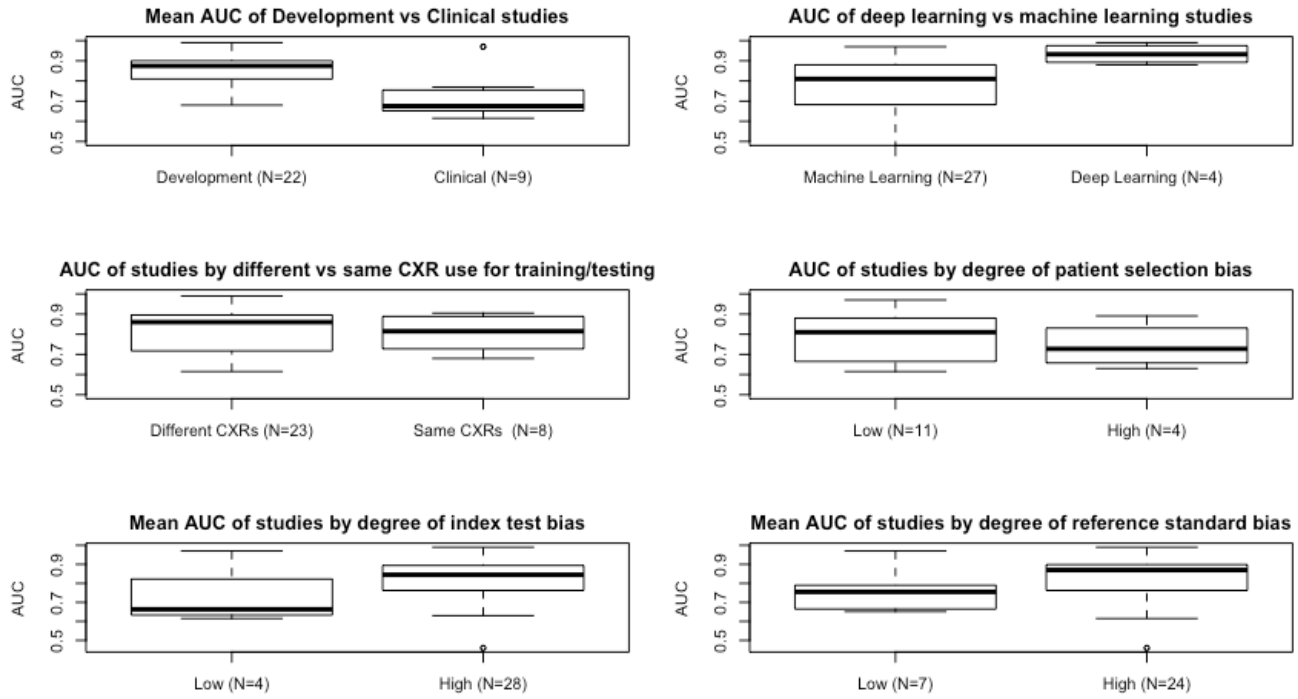

Figure 6. Boxplots of the AUC of studies stratified by software design, CXR usage, reference standard, and degree of patient selection, index test, and reference standard bias

AUC, area under the cure; Vs, versus; CXR, chest x-ray

Recommendations for studies assessing CAD accuracy
• Describe the setting in which CXR were completed: triage vs screening scenario • Apply QUADAS-2 to assess the quality of the databases used • Describe how CXRs were selected for training and testing • Use different CXRs from separate databases for training and testing • Clearly describe how a true positive TB case are defined • Use a microbiologic reference standard • Report and preferably pre-specify a cut-off or threshold score to define a positive or negative if the software has a continuous output • Report how the cut-off or threshold score was determined if reported • State whether pre-training/verification of CAD with local CXRs is required prior to use in each setting • Define in which setting the CAD tool should be used: triage vs screening scenario

Box 1. Recommendations for CAD accuracy study design elements

CHAPTER 5: Additional details and limitations of existing literature

This chapter provides an additional detailed review on the development study databases and their limitations. It also describes CAD4TB in greater detail, the quality of the clinical studies, and highlights the variability of human reading of CXRs for PTB identification. These details could not be succinctly incorporated into a publishable manuscript and are therefore reported here separately.

CAD Databases

As stated previously, to assess the validity of CAD software studies it is important to assess the quality of the databases used to train and test the CAD software and to determine if they were representative of either the triage or screening use-case. The two CXR databases most frequently used in the development studies were the MC and the Shenzhen CXR. The CXR from the former were collected as part of the Montgomery County's Tuberculosis screening program (Maryland, USA) and contains 80 normal cases and 58 abnormal cases with TB manifestations (37). The enrolment strategy is not well described and the use-case (i.e. triage or screening) is unclear. The classification of PTB was made by radiologist reading defined by an abnormality seen in lung consistent with PTB, and were reported as either normal or abnormal (37). The Shenzhen database consists of CXRs from the Shenzhen No.3 People's Hospital (Shenzhen, China), that were taken as part of the daily hospital routine care in September 2012, with consecutive enrolment (37). This dataset set contains 662 frontal CXRS for adult and pediatric patients. Of these, 326 are normal cases and 336 are abnormal with manifestations of PTB (37). These CXRs are also defined as either normal or abnormal, by a human reader (37). However, it was unclear if the patients included had symptoms of TB or had CXRs completed for other purposes.

The other commonly used publicly available data set, the JSRT, was developed from 13 medical centers in Japan and one from the United States (38, 80). This data set does not include PTB cases, but rather is comprised of images with single pulmonary nodules, confirmed by computed tomography and histology. There are 247 CXRs, 154 images with, and 93 images without a pulmonary nodule. Of the 154 CXRs with a pulmonary nodule, 100 were classified as malignant and the remaining 54 classified as benign (38, 80). Another similarly constructed database is the Calgary dataset (68). These datasets have little clinical validity for PTB identification.

The database from two sites in Sub-Saharan Africa collected 945 consecutive digital CXRs for PTB detection purposes (triage use-case) (45, 46) (60, 85). The reference standard used was radiological, with 514 abnormal and 431 normal images. The chest radiograph reading and recording system (CRRS) was used to classify the images. The datasets from Lusaka, Zambia also used CRRS guided human reading as the reference standard and has 917 CXRs (60, 85). The large Zambian dataset, contains 1600 CXRs from consecutively enrolled patients enrolled patients reporting TB symptoms, i.e. representative of the triage scenario (85). The large Zambian data set used a radiologic reference standard. The smaller Zambian data set has 645 CXRS (60), form consecutively enrolled patients enrolled patients also from in the triage use-case setting. The Tanzania and Gambia datasets are other similar clinical databases, with less information available, but the images in the Gambian dataset come from a TB prevalence survey (between December 2011 and January 2013), and therefore represent a screening population (61). The Korean Institute of Tuberculosis (KIT)(86), New Delhi, India dataset (66), and Kenya (87) are not well described.

Four other clinical datasets that were used also had microbiologic reference standards available in addition to radiologic data. These include the Find & Treat dataset, and TB-NEAT database, The Belarus Tuberculosis portal dataset, and the Thomas Jefferson Hospital dataset (TJH) (15, 62, 88). The Find & Treat database is comprised of a screening cohort of high-risk homeless participants, prisoners, and drug users from the United Kingdom. TB cases were based on a

clinical decision to treat TB, and in most cases, this was based on a positive sputum culture (62). The TB-NEAT study evaluated patients from Cape-Town presenting with presumptive TB. For all cases sputum culture results were available (62). The Belarus Tuberculosis portal dataset, and the Thomas Jefferson Hospital dataset (TJH) both had microbiologically confirmed PTB cases, but it was unclear from the papers how the patients in both data sets were included (e.g. consecutively vs case control), and the use-case (triage vs screening) was not specified (15).

CAD4TB Usage

Table 9 summarizes the models of x-ray machines that took the CXRs in each study and the thresholds above which the CXRs were determined to be positive. In the triage use-case the threshold score was pre-specified in 3/6 (50%) of the studies. Of these, one used a pilot study where CXRs from their site were analyzed by CAD4TB to generate a ROC curve which was used to identify a threshold score that optimized sensitivity and specificity which was then prospectively studied. In the screening use-case a pre-specified threshold score was used in only 1/3 (33%) of the studies.

Selection and Enrollment

All triage studies used a consecutive enrollment strategy, with 3/7 (43%) being prospective, 4/7 (57%) retrospective. Breuninger et al (20) completed the analysis on only 66% of cases enrolled. They excluded those lost to follow up (n=270), those with extra-pulmonary TB (n=5), and those diagnosed with symptoms alone with negative cultures and smears (n=134). The remainder of the studies completed their analysis on greater than 90% of those enrolled.

In the screening studies, Melendez et al (43) used data from the Zambia National Tuberculosis Prevalence Survey, which employed a random cluster survey methodology (Table 8). Melendez et al included patients from the Find &Treat study and enrolled 99% of eligible participants. Melendez et al (43) included only those participants from the original TB prevalence survey with complete data, including CXR, clinical, and CAD4TB results, and therefore only 56% of eligible patients were enrolled. 92% of enrolled participants were included in the final analysis

(43). This likely resulted in non-differential selection bias. Muyoyeta et al (82) used data where participants were prospectively consecutively enrolled and Steiner et al (44) conducted a case-control retrospective study (see Chapter 4 Table 8). Demographically speaking, the study populations of all the triage and screening studies with microbiologic references were quite similar in terms of age, HIV prevalence, and TB prevalence (see Chapter 4 Table 4).

CXR reading

As stated previously stated, current WHO guidelines emphasize that CXR alone should not be used for TB diagnosis (7). This is in part due to poor access to expert interpretation, and even when expert interpretation is available there still remains significant interobserver variation (7). This is relevant in CAD studies as many use the specificity or sensitivity of human readers to determine a threshold score for the software. This was true for 5/9 of the CAD4TB studies. Therefore, we examined the type of human readers used in these studies and the reported sensitivities and specificities of the human readers (Table 10).

In the triage use-case 4/6 (67 %) of the studies used human readers, and of these all had at least one expert reader (chest physician, or radiologist). Most only reported expert sensitivity and specificity. However, one study reported accuracies for both non-expert and expert readers and exemplifies the large variability that exists within human interpretation (20). Using liquid culture and/or AFB smear as the reference standard the sensitivity of the non-expert readers was 0.97, while the expert was 0.84, and the specificity of the non-experts was 0.18 while the expert 0.72. Additionally, classification (continuous scores versus categories of abnormalities) and educational qualifications varied across studies.

Within the screening studies 2/3 (67%) used human readers, and of these all had expert readers (radiologists). One study also reported non-expert and expert accuracies (43). A large degree of variation was found: the sensitivity of non-experts was 0.70 while experts had 0.53, and the specificity of non-experts was 0.93 while experts had 1.0 (43). The reference standard used for

these calculations was liquid culture. However, only participants with abnormal (not necessarily suggestive of TB) underwent a culture.

Chapter 5: Figures and Tables

Reference	Model of X-ray machine	CAD threshold score pre-specified	How was CAD threshold score chosen	CAD threshold score used
CAD4TB TRIAGE				
Rahman et al, 2017	Delft EZ DR X-ray system	No	ROC curve was used to identify threshold score that achieved same specificity as field officer	> 62
Melendez et al, 2016	Odelca-DR, Delft Imaging Systems	Yes	Using previous collected CXR data of population to generate ROC curve	> 60
Muyoyeta et al, 2015	Odelca-DR, Delft Imaging Systems	Yes	Using previous collected CXR data of population to generate ROC curve	> 60
Breuninger et al, 2014	Philips Cosmos BS radiography system	No	ROC curve was used to identify threshold score that achieved comparable specificity as field officer	> 55
Muyoyeta et al, 2014	Odelca-DR, Delft Imaging Systems	Yes	Using previous collected CXR data of population to generate ROC curve	> 60
Maduskar et al, 2013	Odelca-DR, Delft Imaging Systems	No	ROC was used to identify a threshold score that achieved comparable sensitivity and specificity to clinical officer score of > 50 (abnormal CXR consistent with TB)	> 50
CAD4TB SCREENING				
Melendez et al, 2017	Easy DR X-ray system; Delft Imaging Systems	No	ROC curve was used to identify threshold score that achieved same specificity as field officer	> 62
Muyoyeta et al, 2017	NR	Yes	Using previous collected CXR data of population to generate ROC curve	> 60
Steiner et al, 2015	Odelca-DR, Delft Imaging Systems	No	ROC curve was used to identify threshold score that achieved same specificity as field officer	> 70

Table 9. Development and Use of CAD4TB threshold score

CAD, computer aided diagnosis; ROC, receiver operator curve; CXR, chest x-ray; TB, tuberculosis

* Patients with a normal CXR by CAD received an AFB smear, while patients with an abnormal CXR as per CAD received NAAT

** Patients with a abnormal CXR as per radiologist reading, or presumptive TB based on TB symptoms received culture/NAAT

Reference	# and type of readers	Expert reader	Non expert readers		Reading x-rays		Ref-erence standard	Non-expert Sn (95% CI)	Non-expert Sp (95% CI)	Expert Sn (95% CI)	Expert Sp (95% CI)
		Qual-ifications (# yrs)	Qual-ifications	Mean # of years	Blinded to micro	Blinded to CAD4TB					
Triage (passive case finding)											
Rahman et al, 2017	1 expert	Radio-logist (> 10 years)	NA	NA	Yes	Yes	NAAT	NA	NA	0.91 (0.89-0.92)	0.58 (0.57-0.59)
Melendez et al, 2016	NR	NA	NA	NA	NA	NA	Liquid culture	NR	NR	NR	NR
Muyoyeta et al, 2015	None	NA	NA	NA	NA	NA	NAAT, AFB	NA	NA	NA	NA
Breuninger et al, 2014	1 expert 1 non-expert	Chest physician (NR)	Clinical officer	NR	Yes	Yes	Liquid culture &/or AFB	0.97 (0.94-0.99)	0.18 (0.13-0.24)	0.84 (0.78-0.89)	0.72 (0.65-0.77)
Muyoyeta et al, 2014	None	NA	NA	NA	NA	NA	NAAT	NA	NA	NA	NA
Maduskar et al, 2013	1 expert, 4 non-experts	Chest radio-logist (NR)	3 year medical diploma	NR	Yes	Yes	Liquid culture &/or AFB	0.83 (0.75 – 0.91)	0.48 (CI NR)	NA	NA
CAD4TB SCREENING active case finding)											
Melendez et al, 2017	4 experts, # number of non-experts NR	Radiologist (> 10 years)	General practitioner	> 2 years	Yes	Yes	Liquid culture †	0.70 (0.60 – 0.78)	0.93 (0.92 – 0.93)	0.53 (0.43 – 0.63) *	1.0 (0.99 –1.0)
Muyoyeta et al, 2017	None	NA	NA	NA	NA	NA	NAAT, AFB*	NA	NA	NA	NA
Steiner et al, 2015	2 experts, 5 non-experts	Radiologist (2 years)	3 Clinical officers, 1 assistant medical officer, 1 less jr Radiologist	NR	Yes	Yes	Radiology	0.63 (CI NR)	0.75 (CI NR)	NA	NA

Table 10. Summary of CAD4TB human readers

#, number; Sn, sensitivity; Sp, specificity; CI, confidence interval; NR, no reported; NA, not applicable; NATT, nucleic acid amplification test; AFB, acid fast bacilli

*NAAT for CAD4TB abnormal, AFB for CAD4TB normal

**Composite reference: positive culture/NAAT and/or initiation of TB treatment

† Only participants with an abnormal CXR as per the non-expert readers underwent culture

CHAPTER 6: Discussion

This review summarized CAD tools in development and those tested clinically for the detection of PTB with CXRs. In assessing the clinical validity of the development studies, we found that the most commonly used datasets, the MC, Shenzhen, and JSRT had either a high degree of patient selection bias, or patient enrollment was unclear. We could not determine if these datasets were representative of either the triage or screening cases. Furthermore, most development studies used the same database for training and testing purposes. This limits the validity of the results.

In terms of clinical usefulness of CAD, none of the development studies, and only 6/10 (60%) of the clinical studies reported pre-specified cut-off scores. Studies that did not use a pre-specified score additionally did not recommend how the results could be applied in the clinical context. Ultimately, CAD will be used to diagnosis PTB. Therefore, microbiologic reference, as the gold standard for diagnosis, should be utilized. The majority of the development studies employed a human reader as their reference standard. As demonstrated in our review, studies that used a human reader as the reference standard reported higher AUC than those that used a microbiologic reference (p-value 0.067). These studies likely systematically overestimate the diagnostic utility of their software by using a human reader, rather than microbiologic reference standard.

The clinical studies evaluated the only currently commercially available software, CAD4TB. The CAD4TB triage studies were demographically quite homogenous. The populations studied had very high HIV and TB prevalence. This limits the generalizability to other patient populations even in high burden areas. Furthermore, while the software achieved high sensitivities with variable specificities for both use-cases across versions, CAD4TB was evaluated using CXRs from datasets that may have also contributed to training the software thereby overestimating the predicative accuracy. Lastly, the software's developers were authors in over half of the published studies introducing potential bias.

This systematic review highlights the need for additional research of CAD CXR reading for PTB identification. Box 1 outlines our recommendations for study design elements for future CAD studies to employ to improve the methodological validity and clinical applicability. Further research that considers the broader health systems in which the test is meant to be used in also needed. The WHO's strategic plan for the use of CXRs in TB care highlights the importance of evaluating existing infrastructure, availability of human resources, and assessing if local stakeholders are interested in using CXRs and CXR based technologies like CAD before recommending widespread uptake and use (12). We conclude that AI based CAD programs are promising, but more clinical studies are needed that minimize sources of potential bias to ensure the validity of the findings.

REFERENCES

1. WILLIAMS FH. The Use of X-Ray Examinations in Pulmonary Tuberculosis. The Boston Medical and Surgical Journal. 1907;157(26):850-3.
2. Pande T, Pai M, Khan FA, Denkinger CM. Use of chest radiography in the 22 highest tuberculosis burden countries. Eur Respir J. 2015;46(6):1816-9.
3. Chunhaswasdikul B, Kamolratanakul P, Jittinandana A, Tangcharoensathien V, Kuptawintu S, Pantumabamrung P. Anti-tuberculosis programs in Thailand: a cost analysis. Southeast Asian J Trop Med Public Health. 1992;23(2):195-9.
4. Samandari T, Bishai D, Luteijn M, Mosimaneotsile B, Motsamai O, Postma M, et al. Costs and consequences of additional chest x-ray in a tuberculosis prevention program in Botswana. Am J Respir Crit Care Med. 2011;183(8):1103-11.
5. Organization WH. Early detection of tuberculosis: an overview of approaches, guidelines and tools. 2011.
6. Organization WH. Frequently asked questions on xpert mtb/rif assay. World Health Organization, Geneva, Switzerland http://www who int/tb/laboratory/xpert_faqs pdf. 2015.
7. WHO. Chest Radiography in Tuberculosis Detection - Summary of current WHO recommendations and guidance on programmatic approaches. 2016 2016.
8. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine. 2018.
9. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-8.
10. WHO. Global Tuberculosis Report 2017. Geneva: Licence: CC BY-NC- SA 3.0 IGO.: World Health Organization; 2017.
11. Pai M, Minion J, Jamieson F, Wolfe J, Behr AM. Canadian Tuberculosis Standards 7th Edition.
12. WHO. Chest radiography in tuberculosis detection: Summary of current WHO recommendations and guidance on programmatic approaches. WHO Library Cataloguing in Publication Data. 2016.
13. The Lung Association TCTS. Canadian Tuberculosis Standards, 7th Edition. 7 ed. Canada: Public Health Agency of Canada; 2014.
14. Doi K. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2007;31(4-5):198-211.
15. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology.284(2):574-82.
16. Hwang S, Kim HE, Jeong J, Kim HJ, editors. A novel approach for tuberculosis screening based on deep convolutional neural networks. Medical Imaging 2016: Computer-Aided Diagnosis; 2016: SPIE.
17. FDA. Guidance for Industry and Food and Drug Administration Staff: Computer-Assisted Detection Devices
Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions. In: Mathematics DoIaA, editor. Rockville, United States2012.

18. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
19. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ open. 2016;6(11):e012799.
20. Breuninger M, Van Ginneken B, Philipsen RHHM, Mhimbira F, Hella JJ, Lwilla F, et al. Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: A validation study from sub-Saharan Africa. PLoS One. 2014;9(9).
21. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126(5):376-80.
22. Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI. Understanding heterogeneity in meta-analysis: the role of meta-regression. Int J Clin Pract. 2009;63(10):1426-34.
23. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review. Int J Tuberc Lung Dis. 2016;20(9):1226-30.
24. Price CP, Christenson RH. Evaluating New Diagnostic Technologies: Perspectives in the UK and US. Clin Chem. 2008;54(9):1421-3.
25. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002;324(7336):539-41.
26. Whiting PF, Rutjes AS, Westwood ME, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
27. Derek R. The Statistical Evaluation of Medical Tests for Classification and Prediction by M. Sullivan Pepe. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(3):656-.
28. How reliable is chest radiography? In: Frieden T, editor. Toman's tuberculosis: case detection, treatment, and monitoring. Questions and answers, second edition. [Internet]. World Health Organization. 2004 [cited 5 August 2018]. Available from: http://apps.who.int/iris/ bitstream/10665/42701/1/9241546034.pdf.
29. Synthesising results when it does not make sense to do a meta-analysis. [Internet]. Cochrane Effective Practice and Organisation of Care (EPOC). 2017 [cited 24/07/2018]. Available from: http://epoc.cochrane.org/resources/epoc-resources-review-authors.
30. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061-6.
31. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ : Canadian Medical Association Journal. 2006;174(4):469-76.
32. Wilczynski NL. Quality of Reporting of Diagnostic Accuracy Studies: No Change Since STARD Statement Publication—Before-and-after Study. Radiology. 2008;248(3):817-23.
33. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1-12.

34. Denkinger CM, Grenier J, Minion J, Pai M. Promise versus Reality: Optimism Bias in Package Inserts for Tuberculosis Diagnostics. J Clin Microbiol. 2012;50(7):2455-61.
35. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: A systematic review. Int J Tuberc Lung Dis. 2016;20(9):1226-30 and ii.
36. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6(7):e1000097.
37. Jaeger S, Candemir S, Antani S, Wang YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery. 2014;4(6):475-7.
38. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol. 2000;174(1):71-4.
39. Abubakar I, Story A, Lipman M, Bothamley G, van Hest R, Andrews N, et al. Diagnostic accuracy of digital chest radiography for pulmonary tuberculosis in a UK urban population. Eur Respir J. 2010;35(3):689-92.
40. Theron G, Zijenah L, Chanda D, Clowes P, Rachow A, Lesosky M, et al. Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial. Lancet. 2014;383(9915):424-35.
41. Muyoyeta M, Moyo M, Kasese N, Ndhlovu M, Milimo D, Mwanza W, et al. Implementation Research to Inform the Use of Xpert MTB/RIF in Primary Health Care Facilities in High TB and HIV Settings in Resource Constrained Settings. PLoS One. 2015;10(6):e0126376.
42. Melendez J, Sanchez CI, Philipsen RH, Maduskar P, Dawson R, Theron G, et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Sci Rep. 2016;6:25265.
43. Melendez J, Philipsen R, C, a-Kapata P, Sunkutu V, Kapata N, et al. Automatic versus human reading of chest X-rays in the Zambia National Tuberculosis Prevalence Survey. International Journal of Tuberculosis & Lung Disease. 2017;21(8):880-6.
44. Steiner A, Mangu C, van den Hombergh J, van Deutekom H, van Ginneken B, Clowes P, et al. Screening for pulmonary tuberculosis in a Tanzanian prison and computer-aided interpretation of chest X-rays. Public health action. 2015;5(4):249-54.
45. Muyoyeta M, Maduskar P, Moyo M, Kasese N, Milimo D, Spooner R, et al. 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. 2014;9(4).
46. Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, Van Ginneken B. Detection of tuberculosis using digital chest radiography: Automated reading vs. interpretation by clinical officers. Int J Tuberc Lung Dis. 2013;17(12):1613-20+i.
47. Fatima A, Akram MU, Akhtar M, Shafique I, editors. Detection of tuberculosis using hybrid features from chest radiographs. SPIE; 2017 2017-1-1. Proceedings of the SPIE, Volume 10225, id. 102252B 5 pp. (2017). SPIE.

48. Ding M, Antani S, Jaeger S, Xue Z, Candemir S, Kohli M, et al., editors. Local-global classifier fusion for screening chest radiographs. SPIE Medical Imaging; 2017 2017-1-1: SPIE.
49. Lopes UK, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology & Medicine. 2017;89:135-43.
50. Udayakumar E, Santhi S, Vetrivelan P. TB screening using SVM and CBC techniques. Current Pediatric Research. 2017;21(2):338-42.
51. Poornimadevi CS, Helen Sulochana C, editors. Automatic detection of pulmonary tuberculosis using image processing techniques. 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016; 2016: Presses Polytechniques Et Universitaires Romandes.
52. Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, et al. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging. 2014;33(2):233-45.
53. Karargyris A, Siegelman J, Tzortzis D, Jaeger S, Candemir S, Xue Z, et al. Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays. Int J Comput Assist Radiol Surg. 2016;11(1):99-106.
54. Jaeger S, Karargyris A, Antani S, Thoma G. Detecting tuberculosis in radiographs using combined lung masks. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:4978-81.
55. Santosh KC, Vajda S, Antani S, Thoma GR. Edge map analysis in chest X-rays for automatic pulmonary abnormality screening. Int J Comput Assist Radiol Surg. 2016;11(9):1637-46.
56. Jaeger S. Detecting Disease in Radiographs with Intuitive Confidence. Sci World J. 2015;2015.
57. Seixas JM, Faria J, Souza Filho JB, Vieira AF, Kritski A, Trajman A. Artificial neural network models to support the diagnosis of pleural tuberculosis in adult patients. Int J Tuberc Lung Dis. 2013;17(5):682-6.
58. Hogeweg L, Sanchez CI, Maduskar P, Philipsen RHHM, Van Ginneken B. Fast and effective quantification of symmetry in medical images for pathology detection: Application to chest radiography. Medical Physics. 2017;44(6):2242-56.
59. Maduskar P, Philipsen RH, Melendez J, Scholten E, Chanda D, Ayles H, et al. Automatic detection of pleural effusion in chest radiographs. Med Image Anal. 2016;28:22-32.
60. Melendez J, van Ginneken B, Maduskar P, Philipsen RH, Ayles H, Sanchez CI. On Combining Multiple-Instance Learning and Active Learning for Computer-Aided Detection of Tuberculosis. IEEE Trans Med Imaging. 2016;35(4):1013-24.
61. Melendez J, Van Ginneken B, Maduskar P, Philipsen RHHM, Reither K, Breuninger M, et al. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays. IEEE Trans Med Imaging. 2015;34(1):179-92.
62. Hogeweg L, Sanchez CI, Maduskar P, Philipsen R, Story A, Dawson R, et al. Automatic Detection of Tuberculosis in Chest Radiographs Using a Combination of Textural, Focal, and Shape Abnormality Analysis. IEEE Trans Med Imaging. 2015;34(12):2429-42.
63. Giacomini G, Miranda JR, Pavan AL, Duarte SB, Ribeiro SM, Pereira PC, et al. Quantification of Pulmonary Inflammatory Processes Using Chest Radiography: Tuberculosis as the Motivating Application. Medicine. 2015;94(26):e1044.

64. Requena-Mendez A, Aldasoro E, Munoz J, Moore DA. Robust and Reproducible Quantification of the Extent of Chest Radiographic Abnormalities (And It's Free!). PLoS One. 2015;10(5):e0128044.
65. Melendez J, Sánchez CI, Philipsen RHHM, Maduskar P, Van Ginneken B, editors. Multiple-instance learning for computer-aided detection of tuberculosis. Medical Imaging 2014: Computer-Aided Diagnosis; 2014; San Diego, CA: SPIE.
66. Chauhan A, Chauhan D, Rout C. Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation. PLoS One. 2014;9(11).
67. Sundaram KM, R, ran CS. An adaptive region growing algorithm with support vector machine classifier for Tuberculosis cavity identification. American Journal of Applied Sciences. 2013;10(12):1616-28.
68. Xu T, Cheng I, M, al M. Automated cavity detection of infectious pulmonary tuberculosis in chest radiographs. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:5178-81.
69. Noor NM, Yunus A, Bakar SA, Hussin A, Rijal OM. Applying a statistical PTB detection procedure to complement the gold standard. Comput Med Imaging Graph. 2011;35(3):186-94.
70. Shen R, Cheng I, Basu A. A hybrid knowledge-guided detection technique for screening of infectious pulmonary tuberculosis from chest radiographs. IEEE transactions on bio-medical engineering. 2010;57(11).
71. Mouton A, Pitcher RD, Douglas TS. Computer-aided detection of pulmonary pathology in pediatric chest radiographs. 2010. p. 619-25.
72. Hogeweg LE, Mol C, Jong PAd, Ginneken Bv, editors. Rib suppression in chest radiographs to improve classification of textural abnormalities2010 2010-1-1: SPIE.
73. Lieberman R, Kwong H, Liu B, Huang H. Computer-assisted detection (CAD) methodology for early detection of response to pharmaceutical therapy in tuberculosis patients. Proceedings of SPIE--the International Society for Optical Engineering. 2009;7260:726030.
74. Arzhaeva Y, Hogeweg L, de Jong PA, Viergever MA, van Ginneken B. Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis. Med Image Comput Comput Assist Interv. 2009;12:724-31.
75. Mohd Noor N, Mohd Rijal O, Shaban H, Ee Ling O. Discrimination between two lung diseases using chest radiographs. Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2005;3:3320-3.
76. Hogeweg LE, Mol C, Jong PAd, Ginneken Bv, editors. Rib suppression in chest radiographs to improve classification of textural abnormalities. SPIE Medical Imaging; 2010 2010-1-1: SPIE.
77. Rahman MT, Codlin AJ, Rahman MM, Nahar A, Reja M, Islam T, et al. An evaluation of automated chest radiography reading software for tuberculosis screening among public- and private-sector patients. Eur Respir J. 2017;49(5).
78. Philipsen RH, Sanchez CI, Maduskar P, Melendez J, Peters-Bax L, Peter JG, et al. Automated chest-radiography as a triage for Xpert testing in resource-constrained settings: a prospective study of diagnostic accuracy and costs. Sci Rep. 2015;5:12215.
79. Muyoyeta M, Kasese NC, Milimo D, Mushanga I, Ndhlovu M, Kapata N, et al. Digital CXR with computer aided diagnosis versus symptom screen to define presumptive tuberculosis among household contacts and impact on tuberculosis diagnosis. BMC Infectious Diseases. 2017;17(1):301.

80. Jaeger S, Karargyris A, C, emir S, Folio L, Siegelman J, et al. Automatic tuberculosis screening using chest radiographs. IEEE Transactions on Medical Imaging. 2014;33(2):233-45.
81. Karargyris A, Siegelman J, Tzortzis D, Jaeger S, C, emir S, et al. Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays. Int J Comput Assist Radiol Surg. 2016;11(1):99-106.
82. Muyoyeta M, Kasese NC, Milimo D, Mushanga I, Ndhlovu M, Kapata N, et al. Digital CXR with computer aided diagnosis versus symptom screen to define presumptive tuberculosis among household contacts and impact on tuberculosis diagnosis. BMC Infect Dis. 2017;17(1):301.
83. Rahman MT, Codlin AJ, Rahman MM, Nahar A, Reja M, Islam T, et al. An evaluation of automated chest radiography reading software for tuberculosis screening among public- and private-sector patients. European Respiratory Journal. 2017;49(5).
84. HARRELL Jr. FE, LEE KL, MARK DB. MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. Stat Med. 1996;15(4):361-87.
85. Maduskar P, Philipsen RH, Melendez J, Scholten E, C, a D, et al. Automatic detection of pleural effusion in chest radiographs. Med Image Anal. 2016;28:22-32.
86. Hwang K, Dang NA, Kuijper S, Gibson T, Anthony R, Claassens MM, et al., editors. A Novel Approach for Tuberculosis Screening Based on Deep Convolutional Neural Networks2011 2011-1-1. RAYYAN-INCLUSION: {"Miriam"=>true, "Amy"=>false} RAYYAN-INCLUSION: {"Miriam"=>false, "Amy"=>true}.
87. Ding M, Antani S, Jaeger S, Xue Z, C, emir S, et al., editors. Local-global classifier fusion for screening chest radiographs2017 2017-1-1. RAYYAN-INCLUSION: {"Miriam"=>true}: SPIE.
88. Arzhaeva Y, Hogeweg L, de Jong PA, Viergever MA, van Ginneken B. Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis. Med Image Comput Comput Assist Interv. 2009;12(Pt 2):724-31.