RESEARCH ARTICLE

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

Miriam Harris[1,2,3]*, **Amy Qi**[2,4], **Luke Jeagal**[4], **Nazi Torabi**[5], **Dick Menzies**[1,4,6], **Alexei Korobitsyn**[7], **Madhukar Pai**[1,4,6], **Ruvandhi R. Nathavitharana**[8], **Faiz Ahmad Khan**[1,4,6]

1 Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada, **2** Department of Medicine, McGill University Health Centre, Montreal, Canada, **3** Department of Medicine, Boston University–Boston Medical Center, Boston, Massachusetts, United States of America, **4** Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute & Research Institute of the McGill University Health Centre, Montreal, Canada, **5** St. Michael's Hospital, Li Ka Shing International Healthcare Education Centre, Toronto, Canada, **6** McGill International TB Centre, Montreal, Canada, **7** Laboratories, Diagnostics & Drug Resistance Global TB Programme WHO, Geneva, Switzerland, **8** Division of Infectious Diseases, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America

* miriam.harris@bmc.org

OPEN ACCESS

Citation: Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, et al. (2019) A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLoS ONE 14(9): e0221339. https://doi.org/10.1371/journal.pone.0221339

Editor: Pascal A. T. Baltzer, Medical University of Vienna, AUSTRIA

Received: February 10, 2019

Accepted: August 5, 2019

Published: September 3, 2019

Data Availability Statement: All relevant data have been deposited in Figshare: https://doi.org/10.6084/m9.figshare.9700166.v1.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

We undertook a systematic review of the diagnostic accuracy of artificial intelligence-based software for identification of radiologic abnormalities (*computer-aided detection*, or CAD) compatible with pulmonary tuberculosis on chest x-rays (CXRs). We searched four databases for articles published between January 2005-February 2019. We summarized data on CAD type, study design, and diagnostic accuracy. We assessed risk of bias with QUADAS-2. We included 53 of the 4712 articles reviewed: 40 focused on CAD design methods ("Development" studies) and 13 focused on evaluation of CAD ("Clinical" studies). Meta-analyses were not performed due to methodological differences. Development studies were more likely to use CXR databases with greater potential for bias as compared to Clinical studies. Areas under the receiver operating characteristic curve (median AUC [IQR]) were significantly higher: in Development studies AUC: 0.88 [0.82–0.90]) versus Clinical studies (0.75 [0.66–0.87]; p-value 0.004); and with deep-learning (0.91 [0.88–0.99]) versus machine-learning (0.82 [0.75–0.89]; $p = 0.001$). We conclude that CAD programs are promising, but the majority of work thus far has been on development rather than clinical evaluation. We provide concrete suggestions on what study design elements should be improved.

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 [1]. In developed countries, chest x-rays (CXRs) have been used for the evaluation of persons presenting with symptoms of possible active pulmonary TB (PTB), and for screening of individuals in high risk groups, for several decades [2]. However, uptake of CXR in high TB burden countries, particularly in resource-constrained settings, has been limited [3, 4].

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 [5]. However, one of the challenges is the paucity of professionals to interpret radiographic images in resource-constrained settings [6]. In recent years, advances in artificial intelligence (AI) technology and methods have led to major progress in automated image recognition by computers. AI has been applied to the analysis of radiologic images to identify abnormalities—referred to as computer-aided detection, or CAD—and represents one potential solution to overcome the personnel shortage. Two commonly used AI approaches that have been used to create CAD programs capable of reading CXRs are 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 [7, 8]. DL is a subset of ML which attempts to model brain architecture [7]. It uses neural networks, or overlaying models, that emphasize learning increasingly meaningful representations of the data [7]. The World Health Organization (WHO) has called for greater evidence before endorsing the use of CAD in PTB diagnostic and screening pathways [5].

To date, there has been only one systematic review of CAD use for PTB detection,[9] and it was limited to reviewing the only commercially available software at the time of publication. Amongst the 5 studies included, the reviewers identified methodological limitations that prevented the 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 [10]. 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 (S1 Appendix). The search strategies included subject headings (where applicable) and text words for the concepts of pulmonary tuberculosis, computer aided diagnosis, and diagnostic accuracy. The search period was limited to papers published after January 1, 2005, and included articles published up to February 13, 2019. 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. Four independent reviewers selected studies for inclusion (MH, AQ, LJ, FAK). Conflicts were reviewed by a third reviewer (FAK).

Data extraction

Data were extracted using a standardized extraction form (S2 Appendix). Three reviewers performed the extraction, with one reviewer (MH) verifying all data forms completed by the second reviewer (AQ & LJ). 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 [5]. 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 [11]. 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 [12–15].

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 [5, 16].

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 software 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. When studies reported more than one AUC, a mean AUC was calculated and used for this analysis. 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 4712 unique citations (Fig 1), of which 2821 studies were excluded at the title and abstract phase. Of the remaining 391, 338 were excluded after full-text review. Amongst the 53 included articles, 40 were classified as Development studies and 13 were classified as Clinical (Table 1). The software developers were either authors or funded the research in 9/13 (69%) of the Clinical studies [17–25], and in 100% (40/40) of the Development studies.

Overview of studies

Within the Development studies, 7/40 (17%) employed DL methods while the remaining 33/40 (83%) used ML approaches (Table 1) [26–65]. 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 [66]. Within the

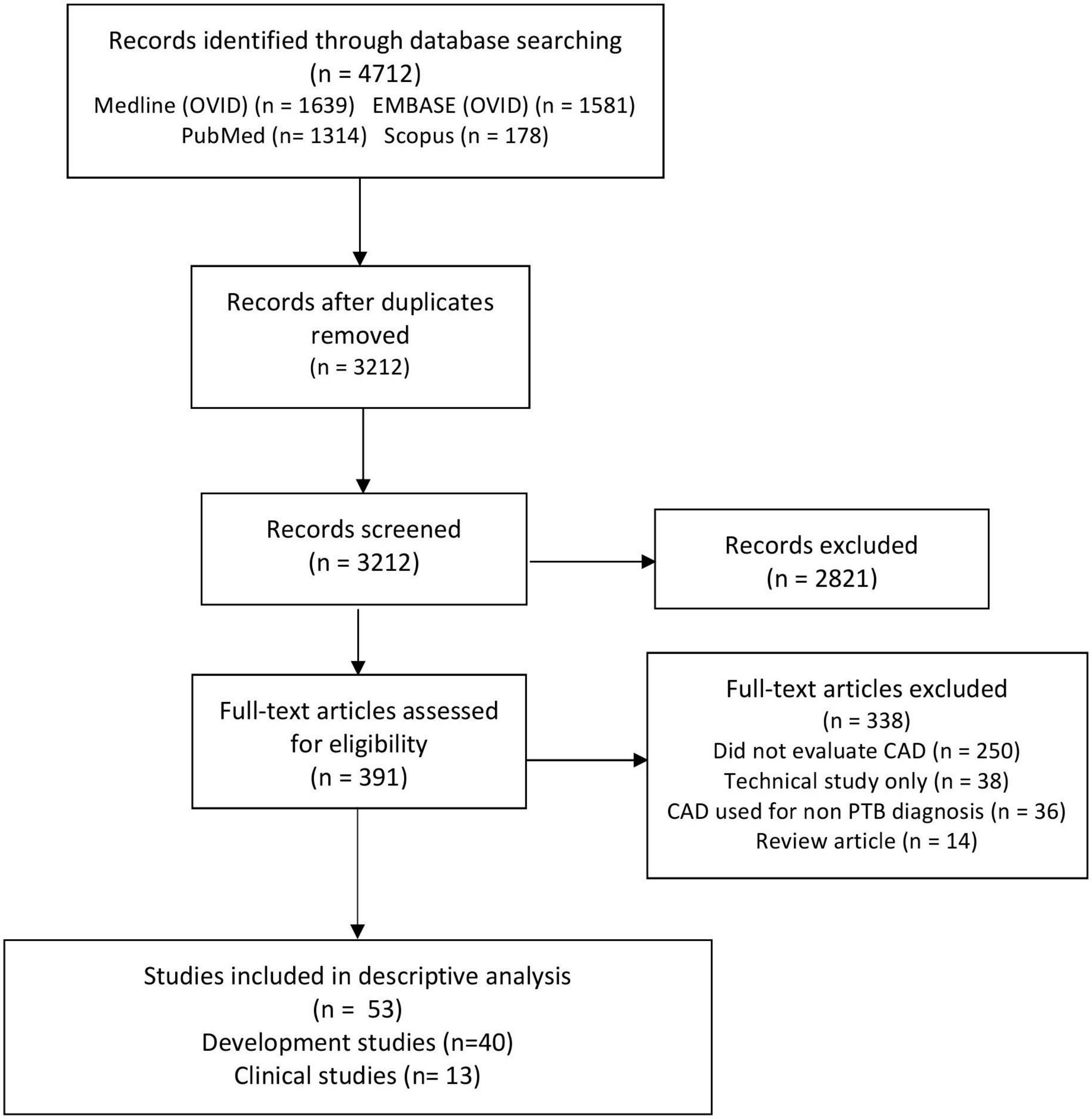

Records identified through database searching
(n = 4712)
Medline (OVID) (n = 1639) EMBASE (OVID) (n = 1581)
PubMed (n= 1314) Scopus (n = 178)
Records after duplicates removed
(n = 3212)
Records screened
(n = 3212)
Records excluded
(n = 2821)
Full-text articles assessed for eligibility
(n = 391)
Full-text articles excluded
(n = 338)
Did not evaluate CAD (n = 250)
Technical study only (n = 38)
CAD used for non PTB diagnosis (n = 36)
Review article (n = 14)
Studies included in descriptive analysis
(n = 53)
Development studies (n=40)
Clinical studies (n= 13)

Fig 1. Study flow diagram. Computer aided detection (CAD).

https://doi.org/10.1371/journal.pone.0221339.g001

Development studies that reported accuracy measures, 3/32 (12%) did not report the database used to train and test their software. Overall, the majority of studies (32/40, 80%), either used the same databases to train and test their software, or did not comment on this (Table 2). For the majority of Development studies demographic data of the study population whose CXR were used to train and evaluate CAD were not reported in detail.

All Clinical studies used ML-based versions of CAD4TB. Within the triage use-case studies, 6/8 (75%) used a microbiologic reference standard on all participants [18, 19, 22, 25, 67, 68]. Within the screening studies, 4/5 (80%) used a microbiologic reference [20, 24, 69, 70]. In two Clinical studies, the CADscore was used to select which participants underwent microbiologic testing, hence the software's diagnostic accuracy could not be assessed [17, 69]. The study

Table 1. Methods of studies included in the descriptive analysis.

Author and year	Country where CXR completed	Databases used	Computer software	Reference standard	Accuracy measures
Development Studies					
Deep learning					
Heo et al, 2019	South Korea	YU AWH	Not named	Human reader	AUC
Hwang et al, 2018	South Korea, USA, China	SNUH, BMC, KUHG, DEMC, MC, CH	DLAD	Liquid culture, NAAT, and or TB treatment	AUC,
Lakhani et al, 2017	USA, China	MC, CH, TJH, Belarus	AlexNet and GoogLeNet	Human reader	AUC, Sn, Sp
Santosh et al, 2017	USA, China, India	MC, CH, IN	Not named	Human reader	AUC, Sn, Sp
Lopes et al, 2017	USA, China	MC, CH	Not named	Human reader	AUC
Santosh et al, 2016	USA, China	MC, CH	Not named	Human reader	AUC
Hwang et al, 2016	South Korea, USA, China	KIT, MC, CH	Alexnet	Human reader	AUC
Machine learning					
Ilena et al, 2018	China	CH	Matlab	Human reader	Sn, Sp, TP, TN, FP, FN
Rajaraman et al, 2018	China, USA, Kenya, India	CH, MC, Kenya, IN	Not named	Human reader	AUC
Sivaramakrishnan et al, 2018	China, USA, Kenya, India	CH, MC, Kenya, IN	Custom 12-layer CNN	Human reader	AUC
Vajda et al, 2018	USA, China	MC, CH	Matlab	Human reader	AUC
Alfadhli et al, 2017	USA	MC	Not named	Human reader	AUC, Sn, TP
Fatima et al, 2017	USA	MC	Not named	Human reader	Sn, Sp
Ding et al, 2017	China, India, Kenya	Kenya, IN, CH	Not named	Human reader	NR
Hogeweg, et al, 2017	Japan, Sub-Saharan Africa	JSRT, Sub-Saharan Africa	Not named	Human reader	AUC
Udayakumar et al. 2017	USA, China	MC, CH	SVM and CBC techniques	Human reader	AUC
Maduskar et al, 2016	Zambia	Large Zambian	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, CH	Not named	Human reader	AUC
Melendez et al, 2016	Zambia	Zambian	Not named	Human reader	AUC
Melendez et al, 2015	Zambia, Tanzania, Gambia	Zambian, Tanzania, Gambian	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	CH	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, CH	Not named	Human reader	AUC, Sn, Sp
Melendez et al, 2014	Zambia, South Africa	Zambian	TB-Xpredict	Human reader	AUC
Chauhan et al, 2014	India	IN	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	Not named	Human reader	Sn, Sp

(*Continued*)

Table 1. (Continued)

Author and year	Country where CXR completed	Databases used	Computer software	Reference standard	Accuracy measures
Shen et al, 2010	Canada	JSRT, Calgary	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, CH	Andrews' curve	Composite reference[**]	NR
Clinical studies					
Machine learning					
Koesoemadinata et al, 2018	Indonesia	Prospective study-specific[†]	CAD4TB (v 5)	Liquid culture/NAAT	AUC, Sn, Sp
Melendez et al, 2018	United Kingdom	Find & Treat	CAD4TB (v 5)	Human reader, TB treatment	AUC, Sn, Sp, TP, FP, TN, FN
Zaidi et al, 2018	Pakistan	Sehatmand Zindagi (Healthy Life)	CAD4TB (v 3.07)	NAAT	AUC, Sn, Sp
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)	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

CXR, chest x-ray; USA, United States of America; UK, United Kingdom; AI, artificial intelligence; YU AWHE, Yonsei University annual worker's health examination; SNUH, Seoul National University Hospital; BMC, Boramae Medical Center; KUHG, Kyunghee University Hospital at Gangdong; DEMC, Daejeon Eulji Medical Center; MC, Montgomery County; CH, Shenzhen Hospital, China; IN, Indian collection New Delhi; TJH, Thomas Jefferson Hospital dataset; JSRT, Japanese Society of Radiology; KIT, Korean Institute of Tuberculosis; F&T, Find and Treat; DLAD, deep learning automatic detection; 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; TN, true negatives, FN, false negatives; ACC, accuracy

[*] Trajman et al. Pleural fluid 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

https://doi.org/10.1371/journal.pone.0221339.t001

Table 2. Accuracy measures reported by development studies.

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									
Heo et al, 2019	YU AWHE	2000	YU AWHE	37475	1202	0.91 (NR), 0.92 (NR)†	NR	NR	NR
Hwang et al, 2018	SNUH	60989	SNUH, BMC, KUHG, DEMC, MC,CH	NR	6768	0.988 (0.976–0.999)	NR	0.95(SNUH), 0.94 (BMC), 1.0 (KUGH), 1.0 (DEMC), 1.0 (MC), 0.95 (CH)*	1.0 (SNUH), 0.96 (BMC), 0.91 (KUGH), 0.98 (DEMC), 0.94 (MC), 0.91 (CH)*
Lakhani et al, 2017	MC,CH, TJH, Belarus	857	MC, CH,TJ, Belarus	150	75	0.99 (0.96–1.00)	NR	0.97 (0.90–1.0)	0.95 (0.87–0.98)
Santosh et al, 2017	MC,CH, IN	976	MC,CH, IN	976	478	0.92 (MC) 0.82 (CH) 0.96 (IN)*	NR	0.88 (MC) 0.78 (CH) 0.92 (IN)*	0.81 (MC) 0.76 (CH) 0.86 (IN)*
Lopes et al, 2017	NR	NR	CHMC, CI,NR	1031	550	0.834 (CH) 0.926 (MC)*	NR	NR	NR
Santosh et al, 2016	NR	NR	CHMC, CI	878	400	0.93 (CH) & 0.88 (MC)*	NR	NR	NR
Hwang et al, 2016	KIT	9221	KIT,MC,CH	2427	NR	0.96*+	NR	NR	NR
Machine learning									
Ilena et al, 2018	CH	20	CH	30	15	NR	NR	0.67 (NR)*	0.86 (NR)*
Rajaraman et al, 2018	CH,MC, AMPATH, Kenya, IN	2073	CH,MC, Kenya, IN	2073	785	0.991 (CH) 0.962 (MC) 0.826 (Kenya) 0.965 (IN)*	NR	NR	NR
Sivaramakrishnan et al, 2018	CH,MC, Kenya, IN	1659	CH,MC, Kenya, IN	1228	785	0.926 (CH), 0.833 (MC), 0.775 (Kenya), 0.956 (IN)*	NR	NR	NR
Vajda et al, 2018	MC,CH	NR	MC,CH	814	392	0.91 (MC), 0.99 (CH)*	NR	NR	NR
Alfadhli et al, 2017	MC	97	MC	41	58	0.89*	NR	0.79*	NR
Fatima et al, 2017	MC	138	MC	138	58	NR	NR	0.83*	0.78*
Udayakumar et al.	MC,CH	NR	MC, CH	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, IN,CH	NR	NR	0.949 (CH), 0.982 (IN), 0.76 (Kenya)*	NR	NR	NR
Maduskar et al, 2016	Large Zambian	629	Large Zambian	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	CH	43	JSRT,CH	NR	NR	0.93*	NR	NR	NR
Melendez et al, 2016	Zambian	461	Zambian	456	248	0.87*	0.45	NR	NR
Melendez et al, 2015	Zambian, Tanzania Gambian	1323	Zambian, Tanzania, Gambian	1313	671	0.86 (Zambia), 0.88 (Tanzania), 0.91 Gambia*	NR	NR	NR

(*Continued*)

Table 2. (Continued)

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)
Hogeweg et al, 2015	F&T, TB-Neat	400	F&T, TB-Neat	400	153	0.87 (0.81–0.92) (F&T), 0.74 (0.69–0.83) (TB-Neat)[#]	NR	NR	NR
Jaeger et al, 2014	MC,CH, JSRT	1000	MC,CH	753	333	0.87[*]	NR	0.78 (0.70–0.85)	0.81 (0.71–0.89)
Melendez et al, 2014	Zambian	461	Zambian	456	NR	0.88[*]	NR	NR	NR
Chauhan et al, 2014	IN	204	IN	102	153	0.96 (0.86–0.99) (DA), 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	60	JSRT, Calgary	60	NR	NR	NR	0.68[*]	0.68[*]
Noor et al, 2011	Retrospective non-clinical	90	Retrospective non-clinical	213	208	NR	NR	0.88[*]	0.84[*]
Shen et al, 2010	JSRT, Calgary	18	JSRT, Calgary	131	19	NR	NR	0.82[*]	NR
Mouton et al, 2010	Clinical non-TB specific	119	Clinical non-TB specific	119	NR	NR	0.78[*]	NR	NR
Hogeweg, et al, 2017	CRASS	348	CRASS, JSRT	498	NR	0.75[*]	NR	NR	NR
Arzhaeva et al, 2009	F&T	217	F&T	217[*++]	37	NR	0.83 TB-sus, 0.74 micro [*†]	NR	NR

CAD, Computer aided detection;; YU AWHE, Yonsei University annual worker's health examination; SNUH, Seoul National University Hospital; BMC, Boramae Medical Center; KUHG, Kyunghee University Hospital at Gangdong; DEMC, Daejeon Eulji Medical Center; MC, Montgomery County; CH, Shenzhen Hospital, China; IN, Indian collection New Delhi; TJH, Thomas Jefferson Hospital dataset; AMPATH, Academic Model Providing Access to Healthcare; JSRT, Japanese Society of Radiology; KIT, Korean Institute of Tuberculosis; 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; TB-sus, TB suspect

[*] 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.

https://doi.org/10.1371/journal.pone.0221339.t002

populations of all the triage studies with microbiologic references were quite similar (S1 and S2 Tables). Notably, the estimated HIV and TB prevalence in the triage studies were quite high, ranging from 15% to 33%. The screening studies had lower TB prevalence compared to triage studies (S1 and S2 Tables).

Quality assessment development studies

We first assessed the databases that were used as sources of CXR images and reference standards for the Development studies (S3 Table). Risk of selection bias was high in 2/18 (11%) of the databases. One dataset did not include PTB cases, and the other only included patients with "typical TB" images [13, 51]. 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 10/18 (56%) studies as a human reader was used, unclear in 3/18 (17%), and low in 4/18 (22%) where a microbiologic reference was used.

The quality of the Development Studies with respect to the assessment of diagnostic accuracy is reported in Fig 2. Selection biased was largely determined by which databases were used (S3 Table). The potential for selection bias was high in 13/33 (39%) studies, unclear in 13/33 (39%), and low in 7/33 (21%). One study [62] had a pre-specified threshold score and therefore had a low risk of bias in the assessment of the index test, but the other 97% had a high risk of bias as the threshold scores were set after the analysis. Additionally, 29/33 (88%) 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 15/33 (45%) studies, in 17/33 (52%) it was unclear, and in 1/33 (3%) it was high.

Quality assessment of clinical studies

All triage studies used a consecutive enrollment strategy, with 3/8 (38%) being prospective, 5/8 (63%) retrospective. Additional details about selection are provided in the Appendix (S2 Table). Fig 3 summarizes the QUADAS-2 assessment of the Clinical studies. There were methodological concerns that likely resulted in a high degree of selection bias in 4/13 (31%) of the studies [18, 21, 23, 68]. This was secondary to case-control design [21], and inappropriate exclusion of patients in the analysis [18, 23, 68]. The threshold score was pre-specified in only 5/13 (38%) of the studies [17, 19, 22, 25, 71]. The remainder of the studies reported threshold scores post-analysis and were therefore determined to have a high risk of bias [18, 20, 21, 23, 24, 68, 70, 72]. The majority of studies, 10/13 (77%) had low potential for bias with regards to the use and performance of the reference standard [18–20, 22–25, 70, 72]. In two studies, the CAD software was used to select patients to undergo microbiologic testing for PTB, and therefore were determined to have a high risk of bias for estimating diagnostic accuracy of CAD [17, 71]. In another study, the reference standard was human reading of the CXR which was deemed to have a high risk of bias [21]. 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 [17, 71], was unclear in 3/10 (30%), and low in 5/10 (50%).

Diagnostic accuracy reported in development studies

We found 33/40 (83%) of the Development studies reported measures of accuracy for index tests. Of the 33 references that did include accuracy assessments, the AUC ranged from 0.78 to 0.99, sensitivity from 0.56 to 0.97, and specificity from 0.36 to 0.95 (Table 2). 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 (Fig 4).

Diagnostic accuracy reported in clinical studies

The forest plots graphically display the diagnostic heterogeneity of the sensitivity and specificity of the triage studies that used a microbiologic reference (Fig 4). In these studies, the sensitivity ranged from 0.86 to 1.00, and specificity ranged from 0.23 to 0.69. In the screening studies, sensitivity ranged from 0.53 to 0.89 and the specificity ranged from 0.56 to 0.98. In one screening study, [21] investigators used a human reader as the reference standard and reported

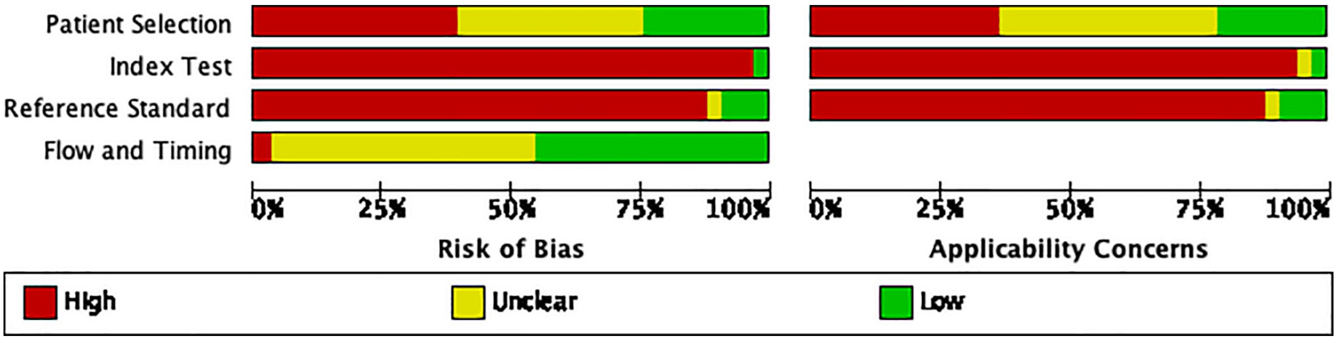


Fig 2. Quality assessment (QUADAS 2) graph of development studies.

https://doi.org/10.1371/journal.pone.0221339.g002

the sensitivity and specificity of CAD were 0.59 and 0.78, respectively. The sensitivity of CAD was higher when using NAAT as the microbiologic reference standard compared to culture. Given the methodological heterogeneity, the lack of standardized threshold scores, and the variability of software versions used, a meta-analysis was not undertaken.

Assessment of study-level factors associated with reported AUC

Fig 5 shows the distribution of reported AUCs stratified by study level characteristics. Reported AUCs were higher in: Development studies (median [IQR] AUC: 0.88 [0.82–0.90]) versus Clinical studies (0.75 [0.66–0.87]; p-value 0.004); and with DL (0.91 [0.88–0.99]) versus ML (0.82 [0.75–0.89]; *p* = 0.001). While not statistically significant, we found that the median AUC of studies using a human reader as the reference standard were higher than those studies using a microbiologic reference standard of 0.88 [0.81–0.90] versus 0.77 [0.67–0.89] respectively (*p* = 0.16). 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 (Fig 5).

Discussion

In this systematic review, we sought to determine the diagnostic accuracy of CAD software programs for detecting PTB on CXRs. Due to study heterogeneity, we did not meta-analyze the data. We identified a number of methodological limitations in the existing evidence base. 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.

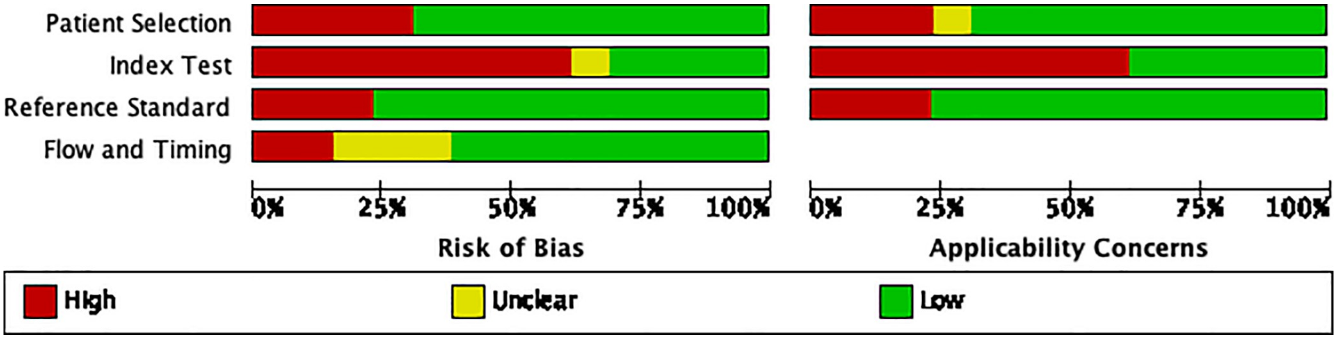
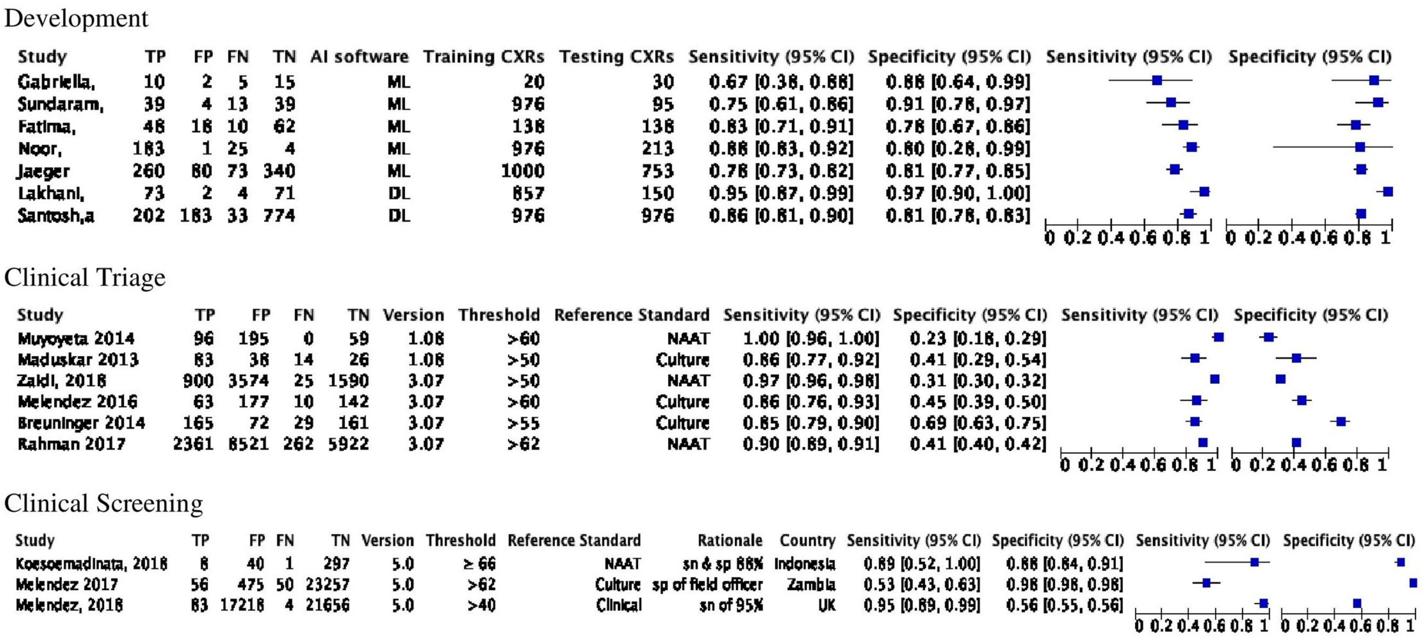


Fig 3. Quality assessment (QUADAS 2) graph of clinical studies.

https://doi.org/10.1371/journal.pone.0221339.g003



Development

Study	TP	FP	FN	TN	AI software	Training CXRs	Testing CXRs	Sensitivity (95% CI)	Specificity (95% CI)
Gabriella,	10	2	5	15	ML	20	30	0.67 [0.38, 0.88]	0.88 [0.64, 0.99]
Sundaram,	39	4	13	39	ML	976	95	0.75 [0.61, 0.86]	0.91 [0.78, 0.97]
Fatima,	48	18	10	62	ML	138	138	0.83 [0.71, 0.91]	0.78 [0.67, 0.86]
Noor,	183	1	25	4	ML	976	213	0.88 [0.83, 0.92]	0.80 [0.28, 0.99]
Jaeger	260	80	73	340	ML	1000	753	0.78 [0.73, 0.82]	0.81 [0.77, 0.85]
Lakhani,	73	2	4	71	DL	857	150	0.95 [0.87, 0.99]	0.97 [0.90, 1.00]
Santosh,a	202	183	33	774	DL	976	976	0.86 [0.81, 0.90]	0.81 [0.78, 0.83]

Clinical Triage

Study	TP	FP	FN	TN	Version	Threshold	Reference Standard	Sensitivity (95% CI)	Specificity (95% CI)
Muyoyeta 2014	96	195	0	59	1.08	>60	NAAT	1.00 [0.96, 1.00]	0.23 [0.18, 0.29]
Maduskar 2013	83	38	14	26	1.08	>50	Culture	0.86 [0.77, 0.92]	0.41 [0.29, 0.54]
Zaidi, 2018	900	3574	25	1590	3.07	>50	NAAT	0.97 [0.96, 0.98]	0.31 [0.30, 0.32]
Melendez 2016	63	177	10	142	3.07	>60	Culture	0.86 [0.76, 0.93]	0.45 [0.39, 0.50]
Breuninger 2014	165	72	29	161	3.07	>55	Culture	0.85 [0.79, 0.90]	0.69 [0.63, 0.75]
Rahman 2017	2361	8521	262	5922	3.07	>62	NAAT	0.90 [0.89, 0.91]	0.41 [0.40, 0.42]

Clinical Screening

Study	TP	FP	FN	TN	Version	Threshold	Reference Standard	Rationale	Country	Sensitivity (95% CI)	Specificity (95% CI)
Koesoemadinata, 2018	8	40	1	297	5.0	≥ 66	NAAT	sn & sp 88%	Indonesia	0.89 [0.52, 1.00]	0.88 [0.84, 0.91]
Melendez 2017	56	475	50	23257	5.0	>62	Culture	sp of field officer	Zambia	0.53 [0.43, 0.63]	0.98 [0.98, 0.98]
Melendez, 2018	83	17218	4	21656	5.0	>40	Clinical	sn of 95%	UK	0.95 [0.89, 0.99]	0.56 [0.55, 0.56]

Fig 4. Forest plots of accuracy measures of development and CAD4TB studies. TP, true positive; FP, false positive; FN, false negative; TN, true negative; AI, artificial intelligence; CXRs, chest x-rays; ML, machine learning; DL, deep learning; CI, confidence interval; NAAT, nucleic acid amplification test.

https://doi.org/10.1371/journal.pone.0221339.g004

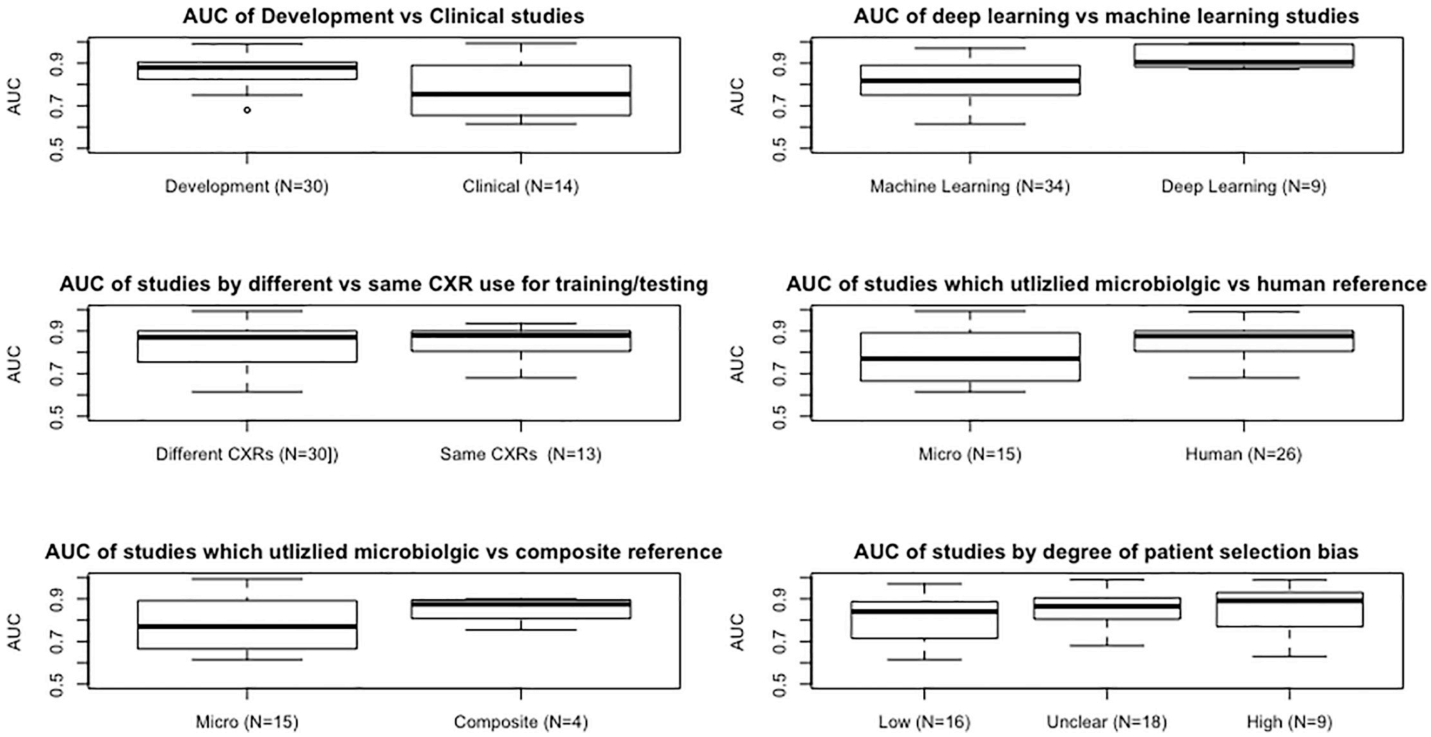


Fig 5. Boxplots of the AUC of studies stratified by software design, CXR usage, reference standard, and degree of patient selection, index test, and reference standard bias. AUC, area under the cure; Vs, versus; CXR, chest x-ray.

https://doi.org/10.1371/journal.pone.0221339.g005

The majority of the CAD evidence base for PTB detection consists of Development studies. While many of these reported some measure of diagnostic accuracy, this was done without assessing the potential risks of bias arising from the databases that were used. 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 the databases used to evaluate CAD in Development studies. We suggest future development studies apply the QUADAS-2 tool to assess for bias of the databases (Box 1).

Box 1. Recommendations for CAD accuracy study design elements

Recommendations for studies assessing CAD accuracy
• For the databases used to assess CAD accuracy, describe whether CXR had been used for triage or screening purposes. • State whether results of the evaluation being reported are applicable to Triage or Screening CXR use-cases
• Apply QUADAS-2 to assess the risk of bias in the databases used to evaluate CAD's diagnostic accuracy
• Describe how CXRs were selected for training and testing
• Use different CXRs from separate databases for training and testing
• Clearly define true positive PTB
• Use a microbiologic reference standard of culture (preferred) or NAAT
• For CAD that output a continuous score, preferably pre-specify the threshold used to differentiate between a positive and negative CAD result.
• For CAD that output a continuous score, report how the threshold score was determined
• State whether pre-training/verification of CAD with local CXRs is required prior to use in each setting

https://doi.org/10.1371/journal.pone.0221339.t003

All Clinical studies evaluated the same commercially available software, CAD4TB. As noted above, meta-analysis was not completed due to the methodological heterogeneity, the lack of standardized threshold scores, and the variability of software versions used. While the software achieved high sensitivities (0.85 to 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, potentially resulting in an overestimation of the predictive power. Lastly, because the populations studied had very high HIV and TB prevalence, the results may 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 (median AUC DL vs ML p-value 0.001), suggesting superior diagnostic accuracy of DL technology. The median AUC of development studies was higher than clinical studies (p-value 0.004). This likely because of the greater risk of bias due to the lack of pre-specified threshold scores, the use of the same databases for training and testing, and the use of a human reader as the reference standard. Our findings also suggested that studies using a human reader reference standard may have systematically overestimated the diagnostic accuracy of CAD, as the median AUC of these studies was higher compared to studies that used a microbiologic reference; the differences were not statistically significant, however. 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 [73].

We suggest some elements that could improve the clinical applicability of future studies of CAD. Studies should include a description of how CXRs were selected for training and testing. Furthermore, CXRs from distinct databases should be used for training and testing. Ideally, accuracy of CAD should be evaluated against a microbiologic reference standard. Lastly, if the software has a continuous output, the threshold score to differentiate between a positive or negative CXR 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 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 and hence unable to comment on the pooled accuracy of CAD.

This systematic review highlights the need for additional research of CAD of PTB on CXR. 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. 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.

Supporting information

S1 Appendix. Search strategies.
(PDF)

S2 Appendix. Extraction form.
(PDF)

S3 Appendix. Prisma (Preferred reporting items for systematic reviews and meta-analyses) checklist.
(DOCX)

S1 Table. Demographics of CAD4TB studies with microbiologic reference standard. CAD, computer aided diagnosis; yrs, years; NR, not reported; TB, tuberculosis; HIV, human immunodeficiency virus
*This is the median, the mean age was not reported.
(PDF)

S2 Table. Selection, enrolment of CAD4TB studies with microbiologic reference standard. NR, not reported; CAD, computer aided diagnosis; NAAT, nucleic acid amplification test
* Patients with an abnormal CXR as per radiologist reading, or presumptive TB based on TB symptoms received culture
** Patients with a normal CXR by CAD received an AFB smear, while patients with an abnormal CXR as per CAD received NAAT.
(PDF)

S3 Table. Quality assessment of datasets used to test and train CAD software of development studies: Risk of bias and applicability concerns. AMPATH, Academic Model Providing Access to Healthcare; CH, Shenzhen Hospital, China; F&T, Find and Treat; IN, Indian collection New Delhi; JSRT, Japanese Society of Radiology; KIT, Korean Institute of Tuberculosis; MC, Montgomery County; YU AWHE, Yonsei University Annual Worker's health examination; SNUH, Seoul National University Hospital; TJH, Thomas Jefferson Hospital dataset; U, unclear; H, high; NA, not applicable; L, low

* Calgary dataset included preselected "typical PTB" images

** 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.

(PDF)

S4 Table. Quality assessment (QUADAS 2) summary of development studies: Risk of bias & applicability concerns.

(PDF)

S5 Table. Quality assessment (QUADAS 2) summary of clinical studies: Risk of bias and applicability concerns.

(PDF)

Acknowledgments

We would like to thank Genevieve Gore from McGill University library services for running an updated search.

Author Contributions

Conceptualization: Miriam Harris, Faiz Ahmad Khan.

Data curation: Miriam Harris, Amy Qi, Nazi Torabi, Faiz Ahmad Khan.

Formal analysis: Miriam Harris, Faiz Ahmad Khan.

Methodology: Faiz Ahmad Khan.

Writing – original draft: Miriam Harris, Faiz Ahmad Khan.

Writing – review & editing: Miriam Harris, Luke Jeagal, Dick Menzies, Alexei Korobitsyn, Madhukar Pai, Ruvandhi R. Nathavitharana, Faiz Ahmad Khan.

References

1. WHO. Global Tuberculosis Report 2017. Geneva: Licence: CC BY-NC- SA 3.0 IGO.: World Health Organization, 2017.
2. WILLIAMS FH. The Use of X-Ray Examinations in Pulmonary Tuberculosis. The Boston Medical and Surgical Journal. 1907; 157(26):850–3. https://doi.org/10.1056/nejm190712261572602
3. 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. https://doi.org/10.1183/13993003.01064-2015 PMID: 26405288
4. 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. Epub 1992/06/01. PMID: 1439970.
5. WHO. Chest Radiography in Tuberculosis Detection—Summary of current WHO recommendations and guidance on programmatic approaches. 2016 2016. Report No.
6. 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. Epub 2010/12/15. https://doi.org/10.1164/rccm.201004-0620OC PMID: 21148723; PubMed Central PMCID: PMC3159079.
7. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018; 319(13):1317–8. https://doi.org/10.1001/jama.2017.18391 PMID: 29532063
8. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine. 2018. https://doi.org/10.1001/jamainternmed.2018.3763 PMID: 30128552

9. 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. https://doi.org/10.5588/ijtld.15.0926 PMID: 27510250

10. 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. https://doi.org/10.1371/journal.pmed.1000097 PMID: 19621072

11. 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. Epub 2011/10/19. https://doi.org/10.7326/0003-4819-155-8-201110180-00009 PMID: 22007046.

12. 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. Epub 2014/12/20. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20 PMID: 25525580; PubMed Central PMCID: PMC4256233.

13. 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. Epub 2000/01/11. https://doi.org/10.2214/ajr.174.1.1740071 PMID: 10628457.

14. 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. Epub 2010/03/02. https://doi.org/10.1183/09031936.00136609 PMID: 20190334.

15. 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. Epub 2013/11/02. https://doi.org/10.1016/S0140-6736(13)62073-5 PMID: 24176144.

16. 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.

17. 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. Epub 2015/06/02. https://doi.org/10.1371/journal.pone.0126376 PMID: 26030301; PubMed Central PMCID: PMC4451006.

18. 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). https://doi.org/10.1371/journal.pone.0106381 PMID: 25192172

19. 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. Epub 2016/04/30. https://doi.org/10.1038/srep25265 PMID: 27126741; PubMed Central PMCID: PMC4850474.

20. 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. https://doi.org/10.5588/ijtld.16.0851 PMID: 28786796.

21. 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. Epub 2016/01/15. https://doi.org/10.5588/pha.15.0037 PMID: 26767179; PubMed Central PMCID: PMC4682617.

22. 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). https://doi.org/10.1371/journal.pone.0093757 PMID: 24705629

23. 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. https://doi.org/10.5588/ijtld.13.0325 PMID: 24200278

24. Melendez J, Hogeweg L, Sanchez CI, Philipsen R, Aldridge RW, Hayward AC, et al. Accuracy of an automated system for tuberculosis detection on chest radiographs in high-risk screening. The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease. 2018; 22(5):567–71. Epub 2018/04/18. https://doi.org/10.5588/ijtld.17.0492 PMID: 29663963; PubMed Central PMCID: PMC5905390.

25. Zaidi SMA, Habib SS, Van Ginneken B, Ferrand RA, Creswell J, Khowaja S, et al. Evaluation of the diagnostic accuracy of Computer-Aided Detection of tuberculosis on Chest radiography among private sector patients in Pakistan. Sci Rep. 2018; 8(1):12339. Epub 2018/08/19. https://doi.org/10.1038/s41598-018-30810-1 PMID: 30120345; PubMed Central PMCID: PMC6098114.

26. 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.

27. 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.

28. Lopes UK, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology & Medicine. 2017; 89:135–43. https://doi.org/10.1016/j.compbiomed.2017.08.001 PMID: 28800442.

29. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology. 284(2):574–82. https://doi.org/10.1148/radiol.2017162326 PMID: 28436741.

30. Udayakumar E, Santhi S, Vetrivelan P. TB screening using SVM and CBC techniques. Current Pediatric Research. 2017; 21(2):338–42. PMID: 617364976.

31. 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.

32. 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.

33. 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. https://doi.org/10.1109/TMI.2013.2284099 PMID: 24108713

34. 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. Epub 2015/06/21. https://doi.org/10.1007/s11548-015-1242-x PMID: 26092662.

35. 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. Epub 2013/02/01. https://doi.org/10.1109/EMBC.2012.6347110 PMID: 23367045.

36. 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. Epub 2016/03/21. https://doi.org/10.1007/s11548-016-1359-6 PMID: 26995600.

37. Jaeger S. Detecting Disease in Radiographs with Intuitive Confidence. Sci World J. 2015;2015. https://doi.org/10.1155/2015/946793 PMID: 26495433

38. 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. Epub 2013/04/12. https://doi.org/10.5588/ijtld.12.0829 PMID: 23575336.

39. 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. https://doi.org/10.1002/mp.12127 PMID: 28134985.

40. 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. Epub 2015/12/22. https://doi.org/10.1016/j.media.2015.09.004 PMID: 26688067.

41. 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. Epub 2015/12/15. https://doi.org/10.1109/TMI.2015.2505672 PMID: 26660889.

42. 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. https://doi.org/10.1109/TMI.2014.2350539 PMID: 25163057

43. 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. Epub 2015/02/24. https://doi.org/10.1109/TMI.2015.2405761 PMID: 25706581.

44. 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. https://doi.org/10.1097/MD.0000000000001044 PMID: 26131814; PubMed Central PMCID: PMC4504622.

45. 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. Epub 2015/05/23. https://doi.org/10.1371/journal.pone.0128044 PMID: 25996917; PubMed Central PMCID: PMC4440724.

46. 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.

47. 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). https://doi.org/10.1371/journal.pone.0112980 PMID: 25390291

48. 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. rayyan-6113211.

49. 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. rayyan-6113466. https://doi.org/10.1109/IEMBS.2011.6091282 PMID: 22255505

50. 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. Epub 2010/11/03. https://doi.org/10.1016/j.compmedimag.2010.10.002 PMID: 21036539.

51. 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).

52. Mouton A, Pitcher RD, Douglas TS. Computer-aided detection of pulmonary pathology in pediatric chest radiographs. 2010. p. 619–25.

53. Hogeweg LE, Mol C, Jong PAd, Ginneken Bv, editors. Rib suppression in chest radiographs to improve classification of textural abnormalities2010 2010-1-1: SPIE.

54. 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. Epub 2009/12/03. https://doi.org/10.1117/12.813583 PMID: 19953192; PubMed Central PMCID: PMC2785044.

55. 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. rayyan-6111190. PMID: 20426176

56. 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. https://doi.org/10.1109/IEMBS.2005.1617187 PMID: 17282956.

57. 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.

58. Alfadhli FHO, Mand AA, Sayeed MS, Sim KS, Al-Shabi M. Classification of tuberculosis with SURF spatial pyramid features. 2017 International Conference on Robotics, Automation and Sciences (ICORAS) 2017. p. 1–5.

59. Gabriella I, Stella A K, Agung W S. Early Detection of Tuberculosis using Chest X-Ray (CXR) with Computer-Aided Diagnosis2018. 76–9 p.

60. Heo SJ, Kim Y, Yun S, Lim SS, Kim J, Nam CM, et al. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. Int J Environ Res Public Health. 2019; 16(2). Epub 2019/01/19. https://doi.org/10.3390/ijerph16020250 PMID: 30654560; PubMed Central PMCID: PMC6352082.

61. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and Validation of a Deep Learning-Based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clin Infect Dis. 2018. Epub 2018/11/13. https://doi.org/10.1093/cid/ciy967 PMID: 30418527.

62. Rajaraman S, Candemir S, Xue Z, Alderson PO, Kohli M, Abuya J, et al. A novel stacked generalization of models for improved TB detection in chest radiographs. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2018; 2018:718–21. Epub 2018/11/18. https://doi.org/10.1109/embc.2018.8512337 PMID: 30440497.

63. Santosh KC, Antani S. Automated Chest X-Ray Screening: Can Lung Region Symmetry Help Detect Pulmonary Abnormalities? IEEE Trans Med Imaging. 2018; 37(5):1168–77. Epub 2018/05/05. https://doi.org/10.1109/TMI.2017.2775636 PMID: 29727280.

64. Sivaramakrishnan R, Petrick N, Mori K, Kohli M, Abuya J, Alderson P, et al. Comparing deep learning models for population screening using chest radiography. Medical Imaging 2018: Computer-Aided Diagnosis2018.

65. Vajda S, Karargyris A, Jaeger S, Santosh KC, Candemir S, Xue Z, et al. Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs. J Med Syst. 2018; 42(8):146. Epub 2018/07/01. https://doi.org/10.1007/s10916-018-0991-9 PMID: 29959539.

66. 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.

67. 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). Epub 2017/05/23. https://doi.org/10.1183/13993003.02159–2016 PMID: 28529202; PubMed Central PMCID: PMC5460641 erj.ersjournals.com.

68. 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. Epub 2015/07/28. https://doi.org/10.1038/srep12215 PMID: 26212560; PubMed Central PMCID: PMC4515744.

69. 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. https://doi.org/10.1186/s12879-017-2388-7 PMID: 28438139.

70. Koesoemadinata RC, Kranzer K, Livia R, Susilawati N, Annisa J, Soetedjo NNM, et al. Computer-assisted chest radiography reading for tuberculosis screening in people living with diabetes mellitus. The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease. 2018; 22(9):1088–94. Epub 2018/08/11. https://doi.org/10.5588/ijtld.17.0827 PMID: 30092877.

71. 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. Epub 2017/04/26. https://doi.org/10.1186/s12879-017-2388-7 PMID: 28438139; PubMed Central PMCID: PMC5402643.

72. 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). https://doi.org/10.1183/13993003.02159-2016 PMID: 28529202.

73. HARRELL FE Jr., 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. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2–4