PLOS GLOBAL PUBLIC HEALTH

RESEARCH ARTICLE

Diagnostic accuracy of three computer-aided detection systems for detecting pulmonary tuberculosis on chest radiography when used for screening: Analysis of an international, multicenter migrants screening study

Sifrash Meseret Gelaw[1]*, Sandra V. Kik[2], Morten Ruhwald[2], Stefano Ongarello[2], Tesfa Semagne Egzertegegne[1], Olga Gorbacheva[3], Christopher Gilpin[3], Nina Marano[4], Scott Lee[4], Christina R. Phares[4], Victoria Medina[1], Bhaskar Amatya[1], Claudia M. Denkinger[2,5]

1 International Organization for Migration (IOM), Manila, Philippines, **2** FIND, Geneva, Switzerland, **3** International Organization for Migration (IOM), Geneva, Switzerland, **4** United States Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, United States of America, **5** Heidelberg University Hospital, Center of Infectious Diseases, Heidelberg, Germany

* sgelaw@iom.int

Check for updates

OPEN ACCESS

Citation: Gelaw SM, Kik SV, Ruhwald M, Ongarello S, Egzertegegne TS, Gorbacheva O, et al. (2023) Diagnostic accuracy of three computer-aided detection systems for detecting pulmonary tuberculosis on chest radiography when used for screening: Analysis of an international, multicenter migrants screening study. PLOS Glob Public Health 3(7): e0000402. https://doi.org/10.1371/journal.pgph.0000402

Editor: Maryam Amour, Muhimbili University of Health and Allied Sciences, UNITED REPUBLIC OF TANZANIA

Received: March 29, 2022

Accepted: June 4, 2023

Published: July 14, 2023

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Data Availability Statement: The data and the chest x-ray images used for this study were obtained from the International Organization for Migration (IOM) pre-migration health assessment (HA) TB screening databases and image archiving system of migrants (refugees and immigrants) bound for the United States. Migrant screening involves multiple stakeholders and complex legal

Abstract

The aim of this study was to independently evaluate the diagnostic accuracy of three artificial intelligence (AI)-based computer aided detection (CAD) systems for detecting pulmonary tuberculosis (TB) on global migrants screening chest x-ray (CXR) cases when compared against both microbiological and radiological reference standards (MRS and RadRS, respectively). Retrospective clinical data and CXR images were collected from the International Organization for Migration (IOM) pre-migration health assessment TB screening global database for US-bound migrants. A total of 2,812 participants were included in the dataset used for analysis against RadRS, of which 1,769 (62.9%) had accompanying microbiological test results and were included against MRS. All CXRs were interpreted by three CAD systems (CAD4TB v6, Lunit INSIGHT v4.9.0, and qXR v2) in offline setting, and re-interpreted by two expert radiologists in a blinded fashion. The performance was evaluated using receiver operating characteristics curve (ROC), estimates of sensitivity and specificity at different CAD thresholds against both microbiological and radiological reference standards (MRS and RadRS, respectively), and was compared with that of the expert radiologists. The area under the curve against MRS was highest for Lunit (0.85; 95% CI 0.83−0.87), followed by qXR (0.75; 95% CI 0.72−0.77) and then CAD4TB (0.71; 95% CI 0.68−0.73). At a set specificity of 70%, Lunit had the highest sensitivity (81.4%; 95% CI 77.9–84.6); at a set sensitivity of 90%, specificity was also highest for Lunit (54.5%; 95% CI 51.7–57.3). The CAD systems performed comparable to the sensitivity (98.3%), and except CAD4TB, to specificity (13.7%) of the expert radiologists. Similar trends were observed when using RadRS. Area under the curve against RadRS was highest for CAD4TB (0.87; 95% CI 0.86–0.89) and Lunit (0.87; 95% CI 0.85–0.88) followed by qXR (0.81; 95% CI

restrictions related to protection of migrants' data. Due to the existing legal conditions as per IN/00138, the Authors will not be able to make publicly available the data nor the CXR images of the data subjects (refugees and immigrants), including anonymized, used in the study. The data is housed at IOM migration health assessment databases subject to the IOM Data Protection Principles (IN/00138). For further information, researchers may contact both of the following parties: • IOM Migration Health Division, Headquarters, Geneva, Switzerland (https://www.iom.int/contact-us) • US Centers for Disease Control and Prevention (CDC), Immigrant, Refugee and Migrant Health Branch, Division of Global Migration and Quarantine, Atlanta, Georgia. (https://www.cdc.gov/contact/index.htm).

Funding: This study has been partially funded through a grant FIND received from an Institute known as Netherlands Enterprise Agency (https://english.rvo.nl/), Reference Number: PDP15CH14. The funder had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript. The funding provided to FIND was an institutional grant for the study and was not granted to specific individuals.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests. SVK, MR, SO, and CMD are or have been employed by FIND. FIND conducts multiple clinical research projects to evaluate new diagnostic tests against published target product profiles that have been defined through consensus processes. These include studies of diagnostic products developed by private sector companies who provide access to know-how, equipment/reagents, and may contribute through unrestricted donations according to FIND policies and in line with guidance from the Organization's external scientific advisory council. This does not alter our adherence to PLOS ONE policies on sharing data and materials. FIND does not attribute any financial value to such access. The other authors have declared that no competing interests exist.

0.80–0.83). At a set specificity of 70%, CAD4TB had highest sensitivity (84.1%; 95% CI 82.3–85.8) followed by Lunit (80.9%; 95% CI 78.9–82.7); and at a set sensitivity of 90%, specificity was also highest for CAD4TB (54.6%; 95% CI 51.3–57.8). In conclusion, the study demonstrated that the three CAD systems had broadly similar diagnostic accuracy with regard to TB screening and comparable accuracy to an expert radiologist against MRS. Compared with different reference standards, Lunit performed better than both qXR and CAD4TB against MRS, and CAD4TB and Lunit better than qXR against RadRS. Moreover, the performance of the CADs can be impacted by characteristics of subgroup of population. The main limitation was that our study relied on retrospective data and MRS was not routinely done in individuals with a low suspicion of TB and a normal CXR. Our findings suggest that CAD systems could be a useful tool for TB screening programs in remote, high TB prevalent places where access to expert radiologists may be limited. However, further large-scale prospective studies are needed to address outstanding questions around the operational performance and technical requirements of the CAD systems.

Introduction

Plain chest radiography remains a crucial tool for early detection of pulmonary tuberculosis (TB) and the monitoring of responses to TB treatment [1]. Chest X-rays (CXRs) have a high sensitivity in detecting pulmonary TB abnormalities, even in asymptomatic TB patients, especially when interpreted by experienced radiologists. Despite this, out of the estimated 10 million global TB cases in 2019, only 7.1 million were detected and reported [2]. As a result, although the global TB incidence rate and annual number of TB deaths has been steadily declining, it is not yet in line with the targets set out in the World Health Organization's (WHO) End TB Strategy [2].

While advances in digital radiography technology have increased the quality of CXR images [3], limited access to these facilities and experienced radiologists remains a long-standing challenge, particularly in low-resource settings with a high TB burden [4]. However, recent advances in artificial intelligence (AI)-based computer-aided detection (CAD) systems have shown promising results in the automated interpretation of CXRs and detection of TB [5–7]. With acceptable accuracy, these CAD systems may help improve access to CXR reading for TB screening and contribute towards achieving WHO's End TB strategy [2, 8]. However, there are limited number of studies in this area, most of which have methodological limitations, studied only one CAD software, with few screening data, and/or industry-funded [7–10]. Moreover, most studies used non-expert CXR interpreters and assessed an online CAD processing system or shared images with the CAD vendors and compared the performance against a suboptimal refence standard of a single sputum specimen tested with Xpert MTB/RIF which further highlights the need for independent and rigorous studies [8–12]. More recent investigations have focused on offline and multiple AI systems [13–15], but they remain few in number.

A global consultation, convened by WHO in 2016, concluded that additional evidence on the performance and use of available CAD systems for TB screening were required [16]. To address this need, the International Organization for Migration (IOM) and FIND entered into a research collaboration to conduct two parallel studies at their respective organizations, both evaluating the accuracy of TB CAD technologies. The studies were conducted independently of the developers, using similar study designs and analysis plans, but involving separate

archives. *The results of both studies have contributed to the updated WHO consolidated guideline TB screening [17].*

Here we present the results of the IOM study, evaluating the diagnostic accuracy of three commercially available CAD systems for detecting TB in offline setting using an independent global archive of CXR images collected from multiple sites performing TB screening of migrants in different countries. against *a microbiological (MRS) and radiological reference standard (RadRS).*

Methods

Ethics statement

The study protocol has received ethical approval from McGill University Health Centre (MUHC) Research Ethics Board (REB) (project approval number 2019–4649). The study has also received IOM legal counsel approval to use the retrospective data and chest x-ray images of participants. CDC approval was obtained in addition for use of data from the pre-migration health assessment program of Migrants bound to the U.S.A. Informed consent was waived by the reviewing institutions since it was not feasible to locate the participants as they already resettled to the receiving country by the time the study was conducted.

Study design

The manufacturer-independent archive of CXR images, set up at the IOM Global Teleradiology and Quality Control Center in Manila, consisted of retrospective clinical data and DICOM CXR images from multiple pre-migration health assessment TB screening of migrants bound for the US. These screenings were conducted across 31 IOM migrant health assessment Centers (MHACs) in 18 different countries between October 2014 and December 2017. The distributions of the countries are summarized in S1 Table for MRS analysis population, and S2 Table for RadRS analysis population. For this study, all CXRs were analyzed by two experienced radiologists, as well as the three CAD systems.

Study participants and screening assessments

The IOM Migration Health Division conducts pre-migration TB screening of refugees and immigrants bound to different resettlement countries through its various MHACs located in different countries worldwide. IOM uses a Global web-based application, Migrant Management Operation System Application (MiMOSA) to record migrants' clinical information during the screening, and Local and Global Picture archive and communication systems (PACS) to archive the CXR images.

The TB screening of US-bound migrants is conducted in accordance with the US Centers for Disease Control and Prevention (CDC) Technical Instructions [18], which includes a clinical history, physical examination, and CXR examination (interpreted by qualified radiologist) using the standardized CXR reporting template provided by the US Department of State, DS-3030. Additionally, if the CXR reading is suggestive of TB or there is a clinical suspicion of TB, three consecutive sputum smear tests, plus solid and liquid culture tests, are completed. Molecular diagnostic tests such as Xpert are also performed if fast results are required or if there is a suspicion of drug-resistance. Participants eligible for inclusion in this study were 15 years or older, with a TB screening CXR for which the initial CXR interpretation and reference standards were available. No images included in this study had ever been shared with any of the CAD system manufacturers.

Sample size and sampling

The sample size was calculated to demonstrate minimum CAD system accuracy targets of 90% sensitivity and 70% specificity, based on the WHO target product profile (TPP) for a TB triage test [19]. The minimum sample size required to detect these sensitivity and specificity targets with 90% power and a 95% confidence interval (CI) of 10% or less was 536 TB and 789 non-TB cases. These numbers were increased by 10% to account for missing information, resulting in a final target population of 590 TB and 868 non-TB cases.

Two samples were drawn from the screening archive (Sample 1 and Sample 2), one for each reference standard. Records were extracted from the MHACs with the highest caseloads first, until the sample size targets had been met.

Data preparations

Biographic information, clinical and laboratory results, and original radiology readings were extracted from the IOM electronic Migrant Management Application, MiMOSA global database, and anonymized before being entered in the study dataset. DICOM CXR images of all participants were collected from each MHACs Local or Global PACS systems, as required, and also anonymized before further use in the study. Clinical and DICOM data were merged into one dataset using a unique participant identifier.

Test methods (index tests and reference standards)

Index tests (CAD systems). At the time of study initiation, the latest versions of commercially available, three "Conformité Européenne" (CE)-marked CAD systems that complied with European Union health, safety, performance, and environmental requirements were installed offline on an IOM-secured server: 1) CAD4TB version 6 (Delft Imaging, Netherlands; henceforth called CAD4TB); 2) Lunit INSIGHT CXR TB algorithm version 4.9.0 (Lunit INSIGHT, South Korea; henceforth called Lunit); and 3) qXR version 2 (Qure.ai, India; henceforth called qXR).

All three CAD systems read posterior-anterior (PA) CXR DICOM images and provide an abnormality score ranging from 0–100 (CAD4TB) or 0–1 (Lunit and qXR). A secondary image with a heatmap (CAD4TB and Lunit) or bounding boxes (qXR) is also produced that indicates the location of the identified abnormal findings (S1 Fig; S1A–S1D Fig). Lunit and qXR have manufacturer-recommended thresholds for TB, while CAD4TB users are required to determine the threshold via a verification process (using data from the user site). Three threshold scores were provided by Lunit, either favoring high sensitivity (score = 0.15), high specificity (score = 0.45) or a middle threshold (score = 0.3). Two thresholds were provided by qXR: a "routine TB screening threshold score" of 0.55 and a "high-risk TB threshold score" of 0.75.

For CAD4TB and qXR, a verification or test run was conducted on sets of CXRs from 13 types of different X-ray machine models used by IOM, which did not form part of this study, as required by the manufacturers at that time. Out of 13 tested, 10 X-ray machine models passed the verification for CAD4TB, and CXRs from those machines were included in the study (Agfa CR10-X, Agfa CR15-X, Agfa CR30-X, CareStream CR975, CareStream DRX-1, CareStream VitaCR, DRGEM, FUJIFILM, Kodak Point Of Care 260, SHIMADZU and SIEMENS). CAD CXR interpretation of DICOM images from study participants was carried out by IOM as per the manufacturer's instructions, using offline server-installed CAD licenses. Only the PA CXR of the initial health assessment for each participant was used for the CAD interpretation, even if some participants had additional CXR views and follow-up CXRs.

Reference standards (microbiological [MRS] and radiological [RadRS]). For MRS analyses, a TB case was defined as a positive result on at least one out of three sputum cultures

collected on consecutive days during the initial screening assessment. A non-TB case was defined as: 1) a negative result for all three sputum cultures, 2) the identification of non-tuberculous mycobacterium; and/or 3) at least one negative sputum culture result if the rest of the samples were contaminated. Only results from specimens taken within 14 days of the CXR were included. In the few cases where Xpert analyses were conducted, a positive Xpert result was interpreted as a positive MRS, even if the culture result was negative.

For RadRS analyses, all CXRs were analyzed by two certified IOM consultant radiologists, with 10 years of experience in TB screening, who had received regular training specific on TB screening CXR interpretation, and who performed best in the regular internal monitoring and evaluation program of IOM using Key performance indicators.

The radiologists were blinded to the clinical and original CXR findings, as well as to the CAD results. Each specialist assessed half of the CXRs using the DS-3030 CXR reporting template containing a specified list of TB and non-TB findings (S3 Table). This re-assessment of the CXRs was conducted to reduce inter-reader variability and standardize the readings, as the original CXR interpretations were performed by several radiologists at different MHACs. If the image quality was not deemed to be acceptable, or if additional CXR views would have been required to complete the interpretation, the radiologists could exclude these CXRs from the analysis. When the new CXR readings showed major discrepancies with the original readings, CXR images were reviewed by a quality control radiologist, who provided a final reading after review of all interpretations from all sources. For RadRS, a TB case was defined as a CXR that was suggestive of active TB disease or old, healed TB (categories 2 and 3 of the CXR classification form; S3 Table). A non-TB case was defined as a normal CXR or one which showed other non-TB findings (categories 1, 4, 5, and 6; S3 Table).

Data analysis

Clinical data, CXR readings, and CAD scores were collated into one dataset and any duplicates identified were excluded prior to analysis. Histograms of the CAD abnormality scores were plotted, receiver operating characteristic (ROC) curves calculated and the area under the curve (AUC) evaluated for each CAD system against both reference standards, using binomial distribution assumptions.

Estimates of sensitivity and specificity were also calculated at: 1) predefined points for sensitivity or specificity based on WHO triage TPP (90% sensitivity and 70% specificity); and 2) manufacturer-provided CAD score thresholds (only for Lunit and qXR) against MRS and RadRS. The sensitivity and specificity of radiologist assessments were also calculated against MRS. Finally, the sensitivity and specificity of each of the CAD systems against MRS were calculated at the threshold that produced the same specificity or sensitivity achieved by the radiologists.

Subgroup analyses for AUC, sensitivity and specificity were conducted for the following groups: age (15–35 years, 36–55 years, 56+ years), sex, geographical region, high-risk groups (e.g., a history of previous TB), migrant type (refugee vs immigrant), HIV status (if known), presence of TB symptoms, sputum smear status, presence of some image quality issues even if the images were deemed acceptable overall, and the presence of additional CXR views obtained during the screening and re-assessed by expert radiologists. Stata software version 16 was used for data management and analysis [20].

Results

Study selection

A total of 2,910 cases were sampled (Fig 1): 589 culture-positive and 865 culture-negative from Sample 1, and 590 CXR suggestive of TB and 866 CXR not suggestive of TB from Sample 2.

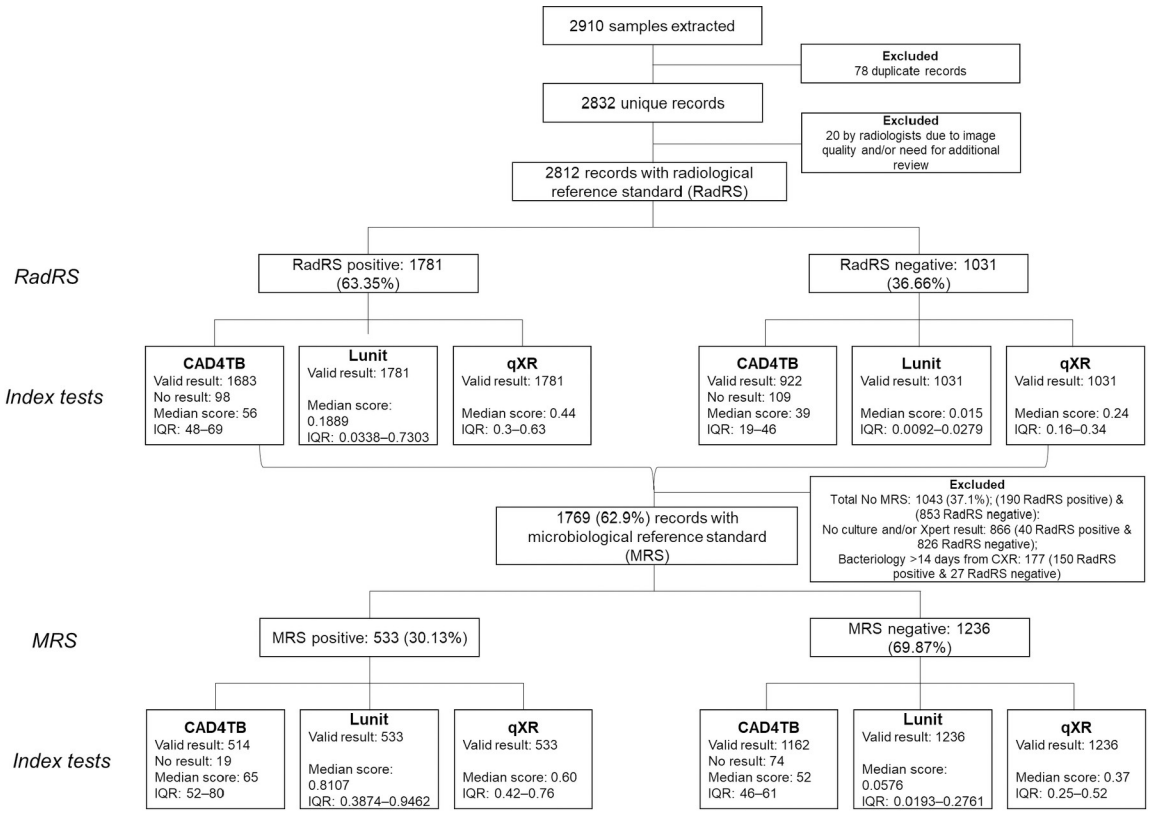


Fig 1. Flowchart of participants included in the analysis.

https://doi.org/10.1371/journal.pgph.0000402.g001

After 78 duplicates were excluded and 20 cases were rejected by the expert radiologists based on the CXRs, 2,812 (1781 RadRS positive and 1031 RadRS negative) cases were included in the RadRS analysis. Of these, 1,769 (533 MRS positive and 1236 MRS negative participants) were also included in the MRS analysis, of which 178 were RadRS positive and 1591 RadRS negative. The remaining 1,043 (190 RadRS positive and 853 RadRS negative) cases were not used for analysis against MRS because either they had > 14 days between the CXR exam and sputum collection, 177 (150 RadRS positive & 27 RadRS negative) or there was no available sputum culture or Xpert result, 866 (40 RadRS positive & 826 RadRS negative). Additionally, 207 participants with invalid CAD4TB scores (empty or negative values) were excluded from analysis for CAD4TB; Lunit and qXR had valid scores for all observations.

Baseline demographic and clinical characteristics

Baseline demographic and clinical characteristics of the whole study population (RadRS) and the population included in the analysis against MRS are presented in S4 Table and Table 1, respectively. Similarly, CXR findings and microbiological test results among RadRS and MRS population are presented in S5 Table and Table 2, respectively.

In the MRS analysis population, more than half (60.5%) were male, most were young (36.2% were 15–35 years of age), 30.1% were MRS positive, and 89.9% had CXR suggestive of TB (RadRS positive). MRS-positive TB cases were reported more often among males (67.0%), at a younger age (44.5% were 15–35 years of age), and in those with TB symptoms compared to non-TB cases (3.6% vs 0.6%). Sputum smears were positive in 29.8% of MRS-positive cases, and abnormal CXR findings in 99%, 97.5% of which were CXR suggestive of TB (Table 1).

Table 1. Baseline demographics and clinical characteristics (MRS analysis population).

Variables	TB (%)	non-TB (%)	Total in MRS (%)
Total	533 (30.1%)	1236 (69.9%)	1769 (100%)
Male	357 (67.0%)	714 (57.8%)	1071 (60.5%)
Female	176 (33.0%)	522 (42.2%)	698 (39.5%)
Age group			
15–35 year	237 (44.5%)	404 (32.7%)	641 (36.2%)
36–55 year	166 (31.1%)	395 (32.0%)	561 (31.7%)
56+ year	130 (24.4%)	437 (35.4%)	567 (32.1%)
Region			
Africa	125 (23.5%)	305 (24.7%)	430 (24.3%)
Asia Pacific	391 (73.4%)	792 (64.1%)	1183 (66.9%)
Middle East	2 (0.4%)	15 (1.2%)	17 (1%)
Eastern Europe	14 (2.6%)	124 (10.0%)	138 (7.8%)
not indicated	1 (0.2%)	0 (0.0%)	1 (0.1%)
TB symptoms			
Cough (any duration)	15 (2.8%)	5 (0.4%)	20 (1.1%)
Fever	5 (0.9%)	1 (0.1%)	6 (0.3%)
Weight loss	11 (2.1%)	3 (0.2%)	14 (0.8%)
Night sweats	5 (0.9%)	0 (0.0%)	5 (0.3%)
TB symptoms present, ≥1	19 (3.6%)	8 (0.6%)	27 (1.5%)
Risk groups			
History of TB	99 (18.6%)	152 (12.3%)	251 (14.2%)
Migrant type (refugee)	325 (61.0%)	692 (56.0%)	1017 (57.5%)
HIV positive	12 (2.3%)	6 (0.5%)	18 (1%)
Image quality issue			
Yes	103 (19.3%)	208 (16.8%)	311 (17.6%)
No	430 (80.7%)	1028 (83.2%)	1458 (82.4%)
Additional view present			
Yes	79 (14.8%)	319 (25.8%)	398 (22.5%)
No	454 (85.2%)	917 (74.2%)	1371 (77.5%)
Smear result			
Positive	159 (30.0%)	16 (1.3%)	175 (9.9%)
Negative	371 (70.0%)	1220 (98.7%)	1591 (89.9%)

https://doi.org/10.1371/journal.pgph.0000402.t001

However, 91% of the MRS negative cases also had abnormal CXR findings, 84.6% of which were CXR suggestive of TB. Only 4.7% of MRS TB cases had Xpert results in addition to cultures, and only two (0.1%) had discrepancies, one being culture-negative and Xpert-positive, and the other being culture-positive and Xpert-negative (Table 2).

In the RadRS analysis population (2812), similarly, most were male (55.3%), younger age group (15–35 year (45.4%), but only 1.2% had one or more TB symptoms and 10% had smear positive result (S4 Table). The 63.3% of RadRS analysis population had CXR findings suggestive of TB (RadRS positive cases. Of those, 32.3% were culture positive (S5 Table).

Histogram distribution of index tests

Abnormality scores of all three CAD systems showed some bimodal distribution when plotted in a two-way histogram against MRS, with a wide range of overlap between TB and non-TB cases (S2A Fig). Of all CAD systems, Lunit provided the least overlap. Similar distributions were observed in the analysis against RadRS (S2B Fig).

Table 2. Chest X-ray findings and microbiological test results (MRS analysis population).

Variables	*TB (%)*	*non-TB (%)*	*Total in MRS (%)*
Total	*533 (30.1%)*	*1236 (69.9%)*	*1769 (100%)*
CXR finding result			
1 = Normal	6 (1.1%)	114 (9.2%)	120 (6.8%)
2 = Abnormal CXR, highly suggestive of active TB (follow-up required)	495 (92.9%)	686 (55.5%)	1181 (66.8%)
3 = Abnormal CXR, may suggest old, healed TB but active TB can't be ruled out (follow-up required)	24 (4.5%)	359 (29.1%)	383 (21.7%)
4 = Abnormal CXR, can remotely suggest old, healed TB but minimal risk of reactivation (NO follow-up required)	3 (0.6%)	53 (4.3%)	56 (3.2%)
5 = Abnormal CXR, not suggestive of TB (follow-up required)	5 (0.9%)	24 (1.9%)	29 (1.6%)
6 = Other abnormalities (NO follow-up required)	0 (0.0%)	0 (0.0%)	0 (0%)
Total	533 (100%)	1236 (100%)	1769 (100%)
Bacteriology result *(for those done = <14 days of CXR to culture date)*			
Culture +, Xpert MTB/RIF not done	450 (84.4%)	0 (0.0%)	450 (25.4%)
Culture +, Xpert MTB/RIF +	81 (15.2%)	0 (0.0%)	81 (4.6%)
Culture +, Xpert MTB/RIF -	1 (0.2%)	0 (0.0%)	1 (0.1%)
Culture -, Xpert MTB/RIF +	1 (0.2%)	0 (0.0%)	1 (0.1%)
Culture -, Xpert MTB/RIF not done	0 (0.0%)	1228 (99.4%)	1228 (69.4%)
Culture -, Xpert MTB/RIF -	0 (0.0%)	8 (0.7%)	8 (0.5%)
Total	533 (100%)	1236 (100%)	1769 (100%)

https://doi.org/10.1371/journal.pgph.0000402.t002

Overall diagnostic accuracy of index tests

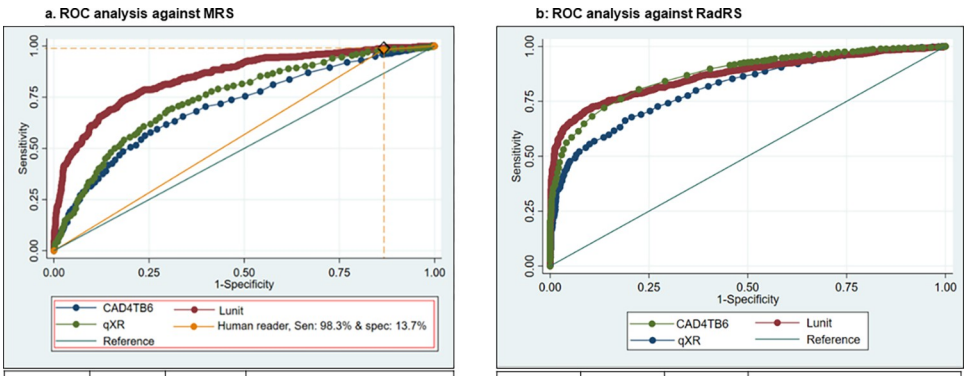
The AUCs of the ROC curves, plotting the estimated sensitivity and specificity at each possible abnormality score against MRS, was highest with the Lunit system (0.85; 95% CI 0.83–0.87), followed by qXR (0.75; 95% CI 0.72–0.77) and then CAD4TB (0.71; 95% CI 0.68–0.73) (Fig 2A).

The ROC curves against RadRS showed greater AUCs for all CAD systems compared with the ROC curves against MRS analysis, being highest with CAD4TB (0.87; 95% CI 0.86–0.89) and Lunit (0.87; 95% CI 0.85–0.88), followed by qXR (0.81; 95% CI 0.80–0.83) (Fig 2B).

The point estimate for the sensitivity (98.3%; 95% CI 96.8–99.2%) and specificity (13.7%; 95% CI 11.8–15.7) of the expert radiologist is presented in the ROC against MRS for visual comparison of the radiologist performance with the performance of the CADs (Fig 2A), and it overlies along the line for Lunit.

Accuracy estimates of index tests

The sensitivity and specificity of the three CAD products computed at a range of thresholds against MRS showed the highest accuracy (a combination of sensitivity and specificity) for Lunit in all target categories, followed by qXR and CAD4TB. At a set sensitivity of 90%, specificity values were 54.5% (95% CI 51.7–57.3%) for Lunit, 32.4% (95% CI 29.8–35.1%) for qXR, and 23.0% (95% CI 20.6–25.5%) for CAD4TB. At a set specificity of 70%, sensitivity values were 81.4% (95% CI 77.9–84.6%) for Lunit, 67.9% (95% CI 63.8–71.9%) for qXR, and 61.7% (95% CI 57.3–65.9%) for CAD4TB.



CAD Product	n	AUC	Confidence Interval
CAD4TB	1,676	0.71	0.68 - 0.73
Lunit	1,676	0.85	0.83 - 0.87
qXR	1,676	0.75	0.72 - 0.77

CAD Product	n	AUC	Confidence Interval
CAD4TB	2,605	0.87	0.86 - 0.89
Lunit	2,605	0.87	0.85 - 0.88
qXR	2,605	0.81	0.80 - 0.83

Fig 2. Receiver operating characteristic curves of three CAD systems against microbiological and radiological reference standards. Cases with no valid CAD4TB scores were excluded from the ROC analysis (n = 93 from the ROC against MRS; n = 207 against RadRS).

https://doi.org/10.1371/journal.pgph.0000402.g002

At the sensitivity value achieved by expert radiologists (98.3% against MRS), the specificities of Lunit and qXR were 15.8% (95% CI 13.8–17.9%) and 12.0% (95% CI 10.2–13.9%), respectively, compared to 13.7% (95% CI 11.8–15.7) for the expert radiologist. CAD4TB had a lower specificity of 6.5% (95% CI 5.2–8.1%). At the specificity value achieved by expert radiologists (13.7% against MRS), the sensitivity was highest for Lunit (99.1%; 95% CI 97.8–99.7%), followed by qXR (97.7%; 95% CI 96.1–98.8%), then CAD4TB (95.9%; 95% CI 93.8–97.5%) (Table 3), compared to 98.3% (95% CI 96.8–99.2%) for the expert radiologist.

At manufacturer-recommended thresholds, the three Lunit thresholds (0.15, 0.3 and 0.45) resulted in sensitivities of 84.2% (95% CI 80.9–87.2), 78.0% (95% CI 74.3–81.5) and 72.6% (95% CI 68.6–76.4), with specificities of 65.5% (95% CI 62.8–68.2), 76.3% (95% CI 73.8–78.6), and 82.6% (95% CI 80.4–84.7) respectively. However, the targets of the WHO TPP for a triage test were not met at any of the thresholds. For qXR, both threshold scores (0.55 and 0.75) resulted in lower sensitivities than Lunit: 56.7% (95% CI 52.3–60.9) and 26.8% (95% CI 23.1–30.8), with specificities of 78.8% (95% CI 76.4–81.1) and 93.0% (95% CI 91.4–94.3) respectively (Table 3).

Sensitivity and specificity estimates against RadRS are presented in S6 Table. At 90% sensitivity, specificity was highest for CAD4TB (54.6%; 95% CI 51.3–57.8), followed by Lunit (45.5%; 95% CI 42.4–48.6) and then qXR (37.5%; 95% CI 34.6–40.6). At 70% specificity, CAD4TB (84.1%; 95% CI 82.3–85.8) and Lunit (80.9%; 95% CI 78.9–82.7) had a higher sensitivity than qXR (70.0%; 95% CI 67.8–72.1) (S6 Table). At the manufacturer-recommended thresholds, both Lunit and qXR showed similar trends to the analysis against MRS, but typically with lower sensitivity and higher specificity values than against MRS (S6 Table). The three Lunit thresholds (0.15, 0.3 and 0.45) against RadRS resulted in sensitivities of 53.1% (95% CI.8–55.5), 42.9% (95% CI 40.6–45.2) and 36.6% (95% CI 34.3–38.8), with specificities of 98.4% (95% CI 97.5–99.1), 99.2% (95% CI 98.5–99.7), and 98.4% (95% CI 97.5–99.1) respectively. For qXR, both threshold scores (0.55 and 0.75) resulted in lower sensitivities than Lunit 35.0% (95% CI 32.8–37.2) and 13.8% (95% CI 12.2–15.4), and specificities of 96.6% (95% CI 95.3–97.6) and 99.7% (95% CI 99.2–99.9) respectively (S6 Table).

Image processing errors were noticed for 178 CXR images after processing by Lunit, in which the images were inverted from the original negative image (bones white) to positive

Table 3. Accuracy estimates of index test at different sensitivity and specificity points (MRS analysis population).

Sensitivity and specificity category	CAD Products	Sensitivity (95% CI)	Specificity (95% CI)	Threshold score
Specificity at 90% sensitivity and associated threshold score	CAD4TB	92.0 (89.3–94.2)	23.0 (20.6–25.5)	46
	Lunit	90.1(87.2–92.5)	54.5 (51.7–57.3)	0.076
	qXR	90.1 (87.2–92.5)	32.4 (29.8–35.1)	0.29
Sensitivity at 70% specificity and associated threshold score	CAD4TB	61.7 (57.3–65.9)	70.9 (67.8–73.1)	59
	Lunit	81.4 (77.9–84.6)	70.1 (67.4–72.6)	0.200
	qXR	67.9 (63.8–71.9)	70.7 (68.1–73.2)	0.49
Specificity at human reader's sensitivity, *98.3% (95% CI 96.8–99.2)*	CAD4TB	98.2 (96.7–99.2)	6.5 (5.2–8.1)	26
	Lunit	98.3 (96.8–99.2)	15.8 (13.8–17.9)	0.013
	qXR	98.1 (96.6–99.1)	12.0 (10.2–13.9)	0.19
Sensitivity at human reader's specificity, *13.7% (95% CI 11.8–15.7)*	CAD4TB	95.9 (93.8–97.5)	13.6 (11.7–15.7)	43
	Lunit	99.1 (97.8–99.7)	13.7 (11.8–15.7)	0.012
	qXR	97.7 (96.1–98.8)	14.2 (12.3–16.3)	0.2
Sensitivity and specificity at manufacturers' recommended threshold score	Lunit, low threshold, high sensitivity	84.2 (80.9–87.2)	65.5 (62.8–68.2)	0.15
	Lunit, middle threshold	78.0 (74.3–81.5)	76.3 (73.8–78.6)	0.3
	Lunit, high threshold, high specificity	72.6 (68.6–76.4)	82.6 (80.4–84.7)	0.45
	qXR, routine TB threshold	56.7 (52.3–60.9)	78.8 (76.4–81.1)	0.55
	qXR, high risk TB threshold	26.8 (23.1–30.8)	93.0 (91.4–94.3)	0.75

The sensitivity of the expert radiologists for detecting TB was 98.3% (95% CI 96.8–99.2) against MRS and the specificity 13.7% (95% CI 11.8–15.7). Lunit scores were provided with 6 decimal places but rounded to three decimals.

https://doi.org/10.1371/journal.pgph.0000402.t003

(bones black). The sensitivity and specificity values, as well as the AUC of Lunit with and without those cases included, were unaffected by these processing errors.

Diagnostic accuracy of index tests in different population subgroups

The diagnostic accuracy, expressed in terms of the AUC of the ROC curve for all three CAD systems, was lower in cases with a history of pulmonary TB compared to those without a history of TB, and lower with CAD4TB in smear-negative cases (0.66; 95% CI 0.63–0.69), compared to smear-positive cases (0.82; 95% CI 0.70–0.94) and in those with additional views (0.58; 95% CI 0.50–0.65), compared to those without additional views (0.72; 95% CI 0.69–0.75) (S3 Fig and S7 Table). For other subgroups such as female sex, HIV infected, absence of TB symptoms, and immigrants (and for Lunit the older age group), CAD systems appeared to show lower AUC estimates compared with their opposing subgroups, though the CIs overlapped. Other subgroups, such as image quality and region, did not show any additional trends (S3 Fig and S7 Table).

Discussion

This study is one of the first comprehensive studies evaluating CAD systems independent of the CAD developers in a population screened for TB using both culture results and expert radiologist assessments as reference standards. The findings from the study demonstrated that the three CAD systems (Lunit, CAD4TB, qXR) have comparable diagnostic accuracy in detecting TB on CXR when used for TB screening and may perform comparably to that of expert radiologists, with Lunit performing better than both qXR and CAD4TB against MRS and CAD4TB and Lunit performing better than qXR against RadRS. However, none of the CAD systems

reached the minimum performance requirements of the WHO triage TPP (90% sensitivity and 70% specificity) [19], in contrast to previously published findings by Khan et al. [13].

The finding that i) Lunit performed best against MRS and ii) CAD4TB performed best against RadRS, shows that the CADs performance can vary by the reference standard used, and may indicate that Lunit is better at detecting CXR findings suggestive of active TB disease, which tend to be culture-positive, while CAD4TB may better detect CXR findings suggestive of old, healed TB that can be identified by radiologists, but tend to be culture-negative. This finding could also reflect the methodology used in the deep machine learning of the CAD product algorithms, e.g., mainly training the software against RadRS versus MRS. The better performance of CAD4TB against RadRS than MRS is also supported by the results of a previous study by Fehr et al. [11].

The low specificity of the CADs at a set sensitivity of 90% against MRS is similar to the expert radiologist and likely is a result of the selection of our study population in whom sputum samples were usually only collected when the initial CXR reading was suggestive of TB or when there was a clinical suspicion of TB. However, also the intrinsic nature of sputum culture, CXR, and CXR signs of TB may be an explanation. Culture analysis detects TB in cases with detectable bacteria in the sputum. As such, it measures the sensitivity of detecting active TB disease, whereas, old, healed pulmonary lesions detected by CXR can be culture-negative and are considered false-positive in the analysis against MRS. Additionally, CXR signs of TB are not specific to TB only, thereby reducing the estimated specificity. Therefore, CXR is recommended for screening but not as a confirmatory diagnostic tool (i.e., a positive CXR TB screening result should be used as criteria for further confirmatory testing, such as sputum cultures, and not for a treatment decision). Nevertheless, for a screening tool the benefit of high sensitivity may outweigh the limitations of a lower specificity. Both Lunit and qXR had relatively lower sensitivity and specificity at all manufacture provided thresholds, though Lunit performed with relatively higher sensitivity while qXR achieved higher specificity. As such, the sensitivity and specificity thresholds of the CADs that correspond with expert radiologist assessments (98.3% and 13.7%, respectively), may be potential candidates for the selection of optimal thresholds for operational use.

Subgroup analyses showed that the performance of CADs can vary among some population demographic and clinical characteristics. All CAD systems performed worse in participants with a history of TB, something which has also been observed in previous studies [13–15]. This is to be expected, as healed TB can leave residual CXR changes, which usually are classified as TB findings on CXR but can lead to negative microbiological test results. CAD4TB performed worse in participants with smear-negative results, in line with the findings of Khan et al. [13]; CAD4TB, moreover, performed worse in cases where additional CXR views were requested by the expert radiologist. Again, these results are not surprising, as smear-positive cases may have obvious CXR abnormalities that can be easily detected, and the absence of a request for additional views may indicate that the initial CXR was of good quality and/or there were no suspicious CXR findings. However, this conclusion was significant only for CAD4TB, while a similar although not statistically significant trend was observed for Lunit and qXR.

Additional trends were observed in the other subgroup analyses. While overlapping CI values indicate that these findings should be interpreted with caution, it appeared that the CAD systems performed less well in females, participants with no TB symptoms, HIV-positive participants, those with an 'immigrant' status compared with those classified with a 'refugee' status, and in older participants (for Lunit only). Other studies have also reported that CAD performance can be significantly impacted by sex and age [13]. The differences in CAD performance among different subgroups indicates that population characteristics should be taken into consideration before implementation.

There are some limitations to this study that should be considered. Firstly, our study relied on retrospective data from a routine migration screening program. As such participants received sputum smear and culture tests during the initial TB screening only when the initial CXR reading was suggestive of TB or there was a clinical suspicion of TB. Therefore, sputum culture testing was not performed for most participants with normal CXRs or CXR findings suggestive of non-TB and were not included in the MRS analysis. This likely resulted in spuriously increase in sensitivity and lower specificity readings for the CADs and expert radiologists, as would be expected from an unselected population. Moreover, as TB cases were overrepresented for both the MRS and RadRS analyses due to the sampling strategy employed in this study, the dataset may not be representative of all people presenting for TB screening but is instead a subset of those who had a higher suspicion of TB and therefore underwent sputum examination. This might have affected the overall accuracy estimates, albeit to a similar extent for all three CAD systems and the expert radiologists, thus, we believe the comparison between the accuracy of CADs versus expert radiologists holds true. Additionally, 20 participants with no radiologist assessment were excluded from the analyses, as well as 207 images from the CAD4TB analysis due to invalid score results. The reason for the invalid scores with the CAD4TB system was unknown, but it could be because the software quality control rejected unacceptable or poor images without requiring further investigation. Although the number of excluded results is small compared with the size of the whole dataset, it is possible that the characteristics of the cases excluded may have been different from those that were included.

Although the study did not systematically evaluate quality control measures of the CADs, some issues were observed during the automated interpretation of the CXRs by the CAD systems. Some CXR projections other than the PA CXRs, such as lateral and lordotic CXR views (which are unsupported by the CADs) or CXRs with image quality issues, were not always flagged by the systems.

The study also did not assess the operational performance of the CADs such as the processing time, technical issues, and troubleshooting responses, infrastructure needs, comparison of offline and online use of the CAD product, cost-effectiveness, or other related matters. In addition, since the study was conducted new versions of the CAD systems have been released and other CAD systems have entered the market [21], which may necessitate further evaluation.

Based on the findings of this study, combined with those of the parallel study conducted by FIND [22], CAD systems may be considered viable as a tool for automated CXR interpretation with regard to TB detection in screening programs, particularly in remote, and/or high TB burden places where there are limited resources and access to expert radiologists. The use of CAD systems in these areas may even have a wider application and contribute to increase the global TB detection rate. Further to these, and other, findings, WHO has recently released consolidated guidelines on tuberculosis recommending that CAD may be used in place of human readers for interpreting digital CXR for TB screening in individuals aged 15 years and older [17]. Another role of CADs, even in places where expert radiologists are available, may be their use for internal quality control monitoring of CXRs complementary to radiologist assessments.

Nevertheless, further studies may be required to investigate the accuracy of CADs in detecting non-TB-significant findings, such as lung cancer or bone lesions as well as the different specific CXR findings suggestive of TB, better address the performance of CADs in the different population subgroups, the way the CADs address image quality issues that might necessitate repeat CXRs or additional views by radiologists, how the CADs handle non-PA CXRs, and non-complied age requirements for specific systems. Likewise, prospective studies are needed to address the operational use of the CAD systems [23], including choice of the CAD system

and version, compatibility with X-ray machine, accepted image format, need for validation, integration into existing workflow and patient registration systems, feasibility of online or off-line use of the software, and technical requirements, as well as the selection of optimal thresholds for the intended use.

In conclusion, the results of this study demonstrated the comparability of the accuracy of three CAD systems for CXR interpretation with regard to TB screening, which may broadly perform similar to that of an expert radiologist. Additionally, the study has demonstrated that the performance of the CAD systems can vary by population demographic and clinical characteristics as well as the reference standard used. As such, these tools may provide viable options for use in TB screening programs to increase TB detection, especially in low resource areas where there may be no available expert radiologists. However, further studies are needed to better address CAD performance in specific population subgroups or different CXR TB findings, to assess other operational and technical factors necessary for proper operational implementation, and to evaluate novel CAD products coming to the market.

Supporting information

S1 Table. Distribution of study participants by across TB screening countries (MRS analysis population).
(XLSX)

S2 Table. Distribution of study participants across TB screening countries (RadRS analysis population).
(XLSX)

S3 Table. Chest X-ray classifications used by IOM radiologists for detecting pulmonary tuberculosis as part of the DS-3030 CXR reporting form.
(XLSX)

S4 Table. Baseline demographics and clinical characteristics (RadRS analysis population).
(XLSX)

S5 Table. Chest X-ray and microbiological test results (RadRS analysis population).
(XLSX)

S6 Table. Estimated diagnostic accuracy of CAD systems (RadRS analysis population).
(XLSX)

S7 Table. Accuracy of the CAD systems in different population subgroups (MRS analysis population).
(XLSX)

S1 Fig. A Sample chest x-ray image before and after the image processing by each CAD system, with image output heat maps/ boxes, S1A–S1D Fig. CAD abnormality scores were 81 for CAD4TB, 94% for Lunit, and 0.86 for qXR. The CXR finding of the case was suggestive of TB and culture positive.
(TIF)

S2 Fig. Two-way histogram distribution of abnormality scores from three CAD systems for a) MRS analysis population, and b) RadRS analysis population.
(TIF)

S3 Fig. Diagnostic accuracy of three CAD systems across population subgroups, S3A-S3G Fig.
(TIF)

Acknowledgments

We would like to thank all the study participants who took part in the IOM health assessment screening and included in the study, the three CAD software companies, Delft Imaging, Qure.ai and Lunit INSIGHT, for installing the CAD products on the IOM server and providing the technical support and troubleshooting, CDC for approving the use of the data and CXR images from the IOM health assessment TB screening of migrants bound to the United States, and Dr. Mary Naughton and Dr. Drew Posey from CDC for assisting in the process, Dr. Cecily MILLER and Dr. Dennis FALZON from the WHO TB programme for their technical collaboration during the study and allowing us to present results of the studies to the WHO TB guideline development group meeting, IOM Legal counsel for approving to use the Data and CXR images of migrants, as well as Dr Fiaz Ahmad Khan from McGill university for his coordination of the Ethical approval. Additionally, we thank our IOM colleague Mr. Rommel Cordero for assisting the technical support and troubleshooting, our IOM Teleradiology Radiologists Dr. Ethel Enriquez-Alas, Dr. Sthennette Jerusalem for re-interpreting the CXRs included in the study, Dr. Lena Maria Ablis-Sun for assisting in quality control review of the discrepant cases and finalizing the CXR report, Ms. May Antonnette Lebanan for her technical guidance during the data analysis, and Dr. Paul Douglas, and Mrs. Jacqueline Weekers and Mr. Enrico Ponziani for the overall support and guidance.

Disclaimer: The findings and conclusions of this article are those of the authors and do not necessarily represent the official position of the US Government or CDC. References in this manuscript to any specific commercial CAD products, process, service, manufacturer, or company do not constitute endorsement or recommendation by the US Government or CDC.

Author Contributions

Conceptualization: Sifrash Meseret Gelaw, Sandra V. Kik, Claudia M. Denkinger.

Data curation: Sifrash Meseret Gelaw, Sandra V. Kik, Stefano Ongarello, Victoria Medina, Bhaskar Amatya.

Formal analysis: Sifrash Meseret Gelaw, Victoria Medina.

Funding acquisition: Sifrash Meseret Gelaw, Sandra V. Kik, Morten Ruhwald, Olga Gorbacheva, Christopher Gilpin, Claudia M. Denkinger.

Investigation: Sifrash Meseret Gelaw, Sandra V. Kik, Claudia M. Denkinger.

Methodology: Sifrash Meseret Gelaw, Sandra V. Kik, Stefano Ongarello, Claudia M. Denkinger.

Project administration: Sifrash Meseret Gelaw, Sandra V. Kik, Morten Ruhwald, Claudia M. Denkinger.

Resources: Sifrash Meseret Gelaw, Sandra V. Kik, Olga Gorbacheva.

Software: Sifrash Meseret Gelaw, Bhaskar Amatya.

Supervision: Sifrash Meseret Gelaw, Sandra V. Kik, Tesfa Semagne Egzertegegne.

Validation: Sifrash Meseret Gelaw, Sandra V. Kik, Morten Ruhwald.

Visualization: Sifrash Meseret Gelaw, Sandra V. Kik, Tesfa Semagne Egzertegegne, Victoria Medina.

Writing – original draft: Sifrash Meseret Gelaw.

Writing – review & editing: Sifrash Meseret Gelaw, Sandra V. Kik, Morten Ruhwald, Stefano Ongarello, Tesfa Semagne Egzertegegne, Olga Gorbacheva, Christopher Gilpin, Nina Marano, Scott Lee, Christina R. Phares, Claudia M. Denkinger.

References

1. World Health Organization. Chest radiography in tuberculosis detection -summary of current WHO recommendations and guidance on programmatic approaches,2016. p. 1–44. Report No.: ISBN 9789241511506. Available from: https://www.who.int/publications/i/item/9789241511506
2. World Health Organization. GLOBAL TUBERCULOSIS REPORT Executive Summary 2020. P. 1–11. ISBN 978-92-4-001313. Available from: https://www.who.int/publications/i/item/9789240013131
3. Bansal GJ. Digital Radiography. A comparison with modern conventional imaging. Postgrad Med J 2006; 82(969):425–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563775/pdf/425.pdf https://doi.org/10.1136/pgmj.2005.038448 PMID: 16822918
4. 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):1814–9. Available from: https://erj.ersjournals.com/content/46/6/1816.
5. Chartrand G, Cheng PM., Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. Radiographics 2017; 37(7):2113–31. Available from: https://doi.org/10.1148/rg.2017170077 PMID: 29131760
6. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017; 284(2):574–82. Available from: https://doi.org/10.1148/radiol.2017162326 PMID: 28436741
7. Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLoS One. 2019; 14(9):e0221339. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221339 PMID: 31479448
8. Ahmad Khan F, Pande T, Tessema B, Song R, Benedetti A, Pai M, et al. Computer-aided reading of tuberculosis chest radiography: moving the research agenda forward to inform policy. Eur Respir J. 2017 Jul; 50(1):1700953. Available from: https://erj.ersjournals.com/content/erj/50/1/1700953.full.pdf https://doi.org/10.1183/13993003.00953-2017 PMID: 28705949
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. Available from: https://www.delft.care/wp-content/uploads/Computer-aided-detection-of-pulmonary-tuberculosis-on-digital.pdf https://doi.org/10.5588/ijtld.15.0926 PMID: 27510250
10. Nash M, Kadavigere R, Andrade J, Sukumar CA, Chawla K, Shenoy VP, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. Sci Rep. 2020; 10(1): 210. Available from: https://doi.org/10.1038/s41598-019-56589-3 PMID: 31937802
11. Fehr J, Konigorski S, Olivier S, Gunda R, Surujdeen A, Gareta D, et al. Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa. NPJ Digit Med., 2021; 4(1):106. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8253848/. https://doi.org/10.1038/s41746-021-00471-y PMID: 34215836
12. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep. 2019; 9(1):15000. Available from: https://www.nature.com/articles/s41598-019-51503-3.pdf. https://doi.org/10.1038/s41598-019-51503-3 PMID: 31628424
13. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. Lancet Digit Health. 2020; 2(11):e573–e581. Available from: https://www.thelancet.com/action/showPdf?pii=S2589-7500%2820%2930221-1. https://doi.org/10.1016/S2589-7500(20)30221-1 PMID: 33328086
14. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel ASS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. Lancet Digit Health. 2021; 3(9):e543–e554Available from: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00116-3/fulltext. https://doi.org/10.1016/S2589-7500(21)00116-3 PMID: 34446265
15. Codlin AJ, Dao TP, Vo LNQ, Forse RJ, Van Truong V, Dang HM, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. Sci Rep. 2021; 11(1):23895. Available

from: https://pubmed.ncbi.nlm.nih.gov/34903808/. https://doi.org/10.1038/s41598-021-03265-0 PMID: 34903808

16. WHO. Chest Radiography in Tuberculosis Detection: Summary of current WHO recommendations and guidance on programmatic approaches 2016, P26–28, ISBN 978 92 4 151150 6 Available from: https://apps.who.int/iris/bitstream/handle/10665/252424/9789241511506-eng.pdf?sequence=1&isAllowed=y
17. WHO. WHO consolidated guidelines on tuberculosis Module 2: Screening–Systematic screening for tuberculosis disease. 2021, P1–68, ISBN: ISBN 978–92–4–002267–6 (electronic version). Available from: https://www.who.int/publications/i/item/9789240022676.
18. The US Centers for Disease Control and Prevention (CDC), Division of Global Migration and Quarantine. Tuberculosis Technical Instructions for Panel Physicians, 2019. Available from: https://www.cdc.gov/immigrantrefugeehealth/panel-physicians/tuberculosis.html
19. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. Geneva, Switzerland; 2014 p. 1–95. Available from: https://www.who.int/tb/publications/tpp_report/en/
20. STATA, the new features in Stata 16. Available from: https://www.stata.com/stata16/
21. Qin ZZ, Naheyan T, Ruhwald M, Denkinger CM, Gelaw S, Nash M, et al. A new resource on artificial intelligence power computer automated detection software products for tuberculosis programmes and implementers. Tuberculosis (Edinb). 2021; 127:102049. Available from: 1-s2.0-S147297922030216X-main.pdf
22. Kik SV, Gelaw SM, Ruhwald M, Song R, Khan FA, Hest RV, et al. Diagnostic accuracy of chest-X-ray reading with three artificial intelligence-based software when used as a screening test for pulmonary tuberculosis: an individual patient meta-analysis of a global chest-x-ray library. Available from: medRxiv 2022.01.24.22269730; [Preprint]. 2022 [posted 2022 January 27]. https://doi.org/10.1101/2022.01.24.22269730.
23. FIND. Digital Chest Radiography and Computer-Aided Detection (CAD) solutions for Tuberculosis Diagnostics Technology Landscape Analysis. 2021. Page 32 to 49. Available from: https://www.researchgate.net/publication/351735151_Digital_Chest_Radiography_and_Computer-Aided_Detection_CAD_Solutions_for_Tuberculosis_Diagnostics_Technology_Landscape_Analysis