Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms

Zhi Zhen Qin, Shahriar Ahmed, Mohammad Shahnewaz Sarker, Kishor Paul, Ahammad Shafiq Sikder Adel, Tasneem Naheyan, Rachael Barrett, Sayera Banu, Jacob Creswell**

Summary

Background Artificial intelligence (AI) algorithms can be trained to recognise tuberculosis-related abnormalities on chest radiographs. Various AI algorithms are available commercially, yet there is little impartial evidence on how their performance compares with each other and with radiologists. We aimed to evaluate five commercial AI algorithms for triaging tuberculosis using a large dataset that had not previously been used to train any AI algorithms.

Methods Individuals aged 15 years or older presenting or referred to three tuberculosis screening centres in Dhaka, Bangladesh, between May 15, 2014, and Oct 4, 2016, were recruited consecutively. Every participant was verbally screened for symptoms and received a digital posterior-anterior chest x-ray and an Xpert MTB/RIF (Xpert) test. All chest x-rays were read independently by a group of three registered radiologists and five commercial AI algorithms: CAD4TB (version 7), InferRead DR (version 2), Lunit INSIGHT CXR (version 4.9.0), JF CXR-1 (version 2), and qXR (version 3). We compared the performance of the AI algorithms with each other, with the radiologists, and with the WHO's Target Product Profile (TPP) of triage tests (≥90% sensitivity and ≥70% specificity). We used a new evaluation framework that simultaneously evaluates sensitivity, proportion of Xpert tests avoided, and number needed to test to inform implementers' choice of software and selection of threshold abnormality scores.

Findings Chest x-rays from 23 954 individuals were included in the analysis. All five AI algorithms significantly outperformed the radiologists. The areas under the receiver operating characteristic curve were 90·81% (95% CI 90·33–91·29) for qXR, 90·34% (89·81–90·87) for CAD4TB, 88·61% (88·03–89·20) for Lunit INSIGHT CXR, 84·90% (84·27–85·54) for InferRead DR, and 84·89% (84·26–85·53) for JF CXR-1. Only qXR (74·3% specificity [95% CI 73·3–74·9]) and CAD4TB (72·9% specificity [72·3–73·5]) met the TPP at 90% sensitivity. All five AI algorithms reduced the number of Xpert tests required by 50% while maintaining a sensitivity above 90%. All AI algorithms performed worse among older age groups (>60 years) and people with a history of tuberculosis.

Interpretation AI algorithms can be highly accurate and useful triage tools for tuberculosis detection in high-burden regions, and outperform human readers.

Funding Government of Canada.

Lancet Digit Health 2021; 3: e543–54

See **Comment** page e535

For the Bengali translation of the abstract see **Online** for appendix 1

For the French translation of the abstract see **Online** for appendix 2

*Senior authors

Stop TB Partnership, Geneva, Switzerland (Z Z Qin MSc, T Naheyan BS, R Barrett BA, J Creswell PhD); **International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, Bangladesh** (S Ahmed MHE, M S Sarker BSc, K Paul MPH, A S S Adel MPH, S Banu PhD)

Correspondence to: Zhi Zhen Qin, Stop TB Partnership, 1218 Le Grand-Saconnex, Geneva, Switzerland **zhizhenq@stoptb.org**

Introduction

The use of artificial intelligence (AI) technology for medical diagnostics has accelerated rapidly in the past decade and AI-powered deep learning neural networks are increasingly being used to analyse medical images, such as chest radiographs or x-rays.[1,2]

Chest x-ray is recommended by WHO as a screening and triage tool for tuberculosis,[3] a disease which killed almost as many people worldwide in 2020 as COVID-19.[4] A triage test is used among people with tuberculosis symptoms or key risk factors for tuberculosis.[5] The performance of chest x-ray as a screening and triage tool has been limited by high inter-reader and intra-reader variability and moderate specificity,[3] as well as limited radiologist availability, especially in countries with a high burden of tuberculosis. Bangladesh is one such country, with tuberculosis prevalence estimated at 260 cases per 100 000 population and with greater prevalence in urban areas.[6]

AI technologies provide an opportunity to vastly increase image reading capacity in a variety of contexts. Such technology makes use of neural networks and deep learning to identify tuberculosis-related abnormalities from chest x-rays.[2] Inspired by the human nervous system, neural networks are interconnected functions, each comprised of a weight and a bias coefficient.[7] Through back-propagation, the networks learn by adjusting the weights and biases of the underlying functions on the basis of the difference between predictions and ground truth in a training dataset.[7] Deep neural networks are

Research in context

Evidence before this study

We searched PubMed on Nov 25, 2020, using the search term ("Tuberculosis" [MeSH] OR "tuberculosis" [tiab]) AND ("artificial intelligence" [tiab] OR "CAD4TB" [tiab] OR "computer-aided interpretation" [tiab] OR "computer aided detection" [tiab] OR "deep learning" [tiab] OR "convolutional neural networks" [tiab] OR "machine learning" [tiab] OR "automatic" [tiab] OR "computer-aided reading" [tiab] OR "automated" [tiab] OR "chest radiographs" [tiab]). Of the 1927 results, only studies evaluating the performance of commercially available artificial intelligence algorithms reading chest x-rays for tuberculosis against an Xpert MTB/RIF reference standard were considered. Development studies, literature reviews, and evaluations against other reference standards (radiological, smear, or culture) were excluded as their findings are not comparable to this study. Five studies were identified that met these criteria, all of which used area under the receiver operating characteristic curve to evaluate the accuracy of CAD4TB, with one publication also including Lunit INSIGHT and qXR. Risk of bias was also high, as three of these studies had authors with commercial interest in the software. Accuracy was generally high across studies, but studies were mixed on how performance compared with human readers.

Added value of this study

Where quality impartial research on this topic is scarce, this study offers an unbiased evaluation using a large dataset, evaluating InferRead DR, JF CXR-1, and the most recent versions of CAD4TB and Lunit INSIGHT for the first time. We also show that area under the precision–recall curve is a better indication of product accuracy in datasets with low disease prevalence than the metrics used in previous publications. We used a new analytical framework assessing implementation-relevant parameters, such as cost-effectiveness, to highlight differences in product performance not observable when solely reporting accuracy.

Implications of all the available evidence

This study shows that computer-aided detection can outperform humans when reading chest x-rays for tuberculosis in a triage setting in a country with high burden of tuberculosis. This finding carries substantial implications for international policy and practice regarding the use of these tools to triage for tuberculosis. The new methods of evaluation suggested in this study should be considered in future studies, as evaluation frameworks more attuned to the needs of implementers could yield literature more relevant to implementers.

structured in a number of layers, which increases the capacity of the machine to perform complex processes, such as parsing medical images.[8]

See Online for appendix 3

Several commercial AI algorithms have emerged in recent years promising to identify tuberculosis-related abnormalities from digital chest x-ray images.[9] AI algorithms produce a continuous abnormality score (from 0 to 100 or from 0 to 1) that represents the likelihood of the presence of tuberculosis-associated abnormalities.[10] Although some software comes with preset threshold abnormality scores, all algorithms also allow users to customise the threshold score at any level to dichotomise the output into binary classifications (either suggesting confirmatory testing for tuberculosis or not).[10]

In March, 2021, WHO updated their tuberculosis screening guidelines to recommend computer-aided detection software in place of human readers for analysis of digital chest x-ray for tuberculosis screening and triage in individuals older than 15 years.[11] WHO did not recommend specific products, leaving many considerations before a decision whether to implement an AI algorithm—and if so, which algorithm—can be made. Most available publications on AI algorithms feature earlier versions of one algorithm and were done with the involvement of the developers. Published evidence from impartial authors is therefore scarce.[12–16] There is also a lack of sizeable external datasets to directly compare algorithms.[15,17] Furthermore, national tuberculosis programmes and health professionals need performance measurements beyond accuracy, which is commonly reported as area under the receiver operating characteristic (ROC)[18] curve (AUROC; appendix 3 p 2),[15] as well as guidance on operating point selection for different patient sources. To help implementers assess accuracy, we evaluated five AI algorithms for triaging tuberculosis using a large dataset that had not previously been used to train any commercial AI algorithms. We also present a new analytical framework for selecting AI software and threshold abnormality scores in different settings.

Methods

Study setting and population

This evaluation of commercially available AI algorithms to read chest x-ray for tuberculosis followed the Standards for Reporting of Diagnostic Accuracy Initiative on design and conduct of diagnostic accuracy evaluations.[19] In this retrospective study, we included all individuals aged 15 years or older who presented or were referred to three tuberculosis screening centres in Dhaka, Bangladesh, between May 15, 2014, and Oct 4, 2016 (appendix 3 p 3).[20] Younger individuals were not included in the analysis as some of the AI algorithms assessed are only approved for use in individuals aged 15 years or older (appendix 3 pp 4–6).

All enrolled participants provided informed written consent (appendix 3 p 3). The study protocol was reviewed and approved by the Research Review Committee and the Ethical Review Committee at the International Centre for Diarrheal Disease Research, Bangladesh (protocol PR-13003).

Reading and testing process

Each participant was verbally screened by health-care workers for tuberculosis symptoms (appendix 3 p 3) using a standardised digital questionnaire and received a digital posterior-anterior chest x-ray from a stationary Delft Easy DR X-ray System (see appendix 3 p 7 for machine specification and radiologist reading details). Asymptomatic people who were referred with suspected tuberculosis by their physicians also received a chest x-ray. Three radiologists, registered with the Bangladesh Medical and Dental Council, who had 10 years, 6 years, and 1 year of experience, respectively (each doing a minimum of 10 000 chest x-ray reads a year), worked part time for this project and took turns in reading chest x-rays. The radiologists read 15–20 chest x-rays per day and were blinded to any information except age and sex, which were present in the metadata of the chest x-ray DICOM files. They graded each chest x-ray as normal or abnormal according to the tuberculosis prevalence survey handbook,[21] and further classified abnormal chest x-rays into three categories to be analysed separately: highly suggestive of tuberculosis, possibly tuberculosis, and abnormal but not tuberculosis.

The following AI companies agreed to participate in this independent study: CAD4TB (version 7) by Delft Imaging Systems (Netherlands), InferRead DR (version 2) by Infervision (China), Lunit INSIGHT CXR for Chest Radiography (version 4.9.0) by Lunit (South Korea), JF CXR-1 (version 2) by JF Healthcare (China), and qXR (version 3) by Qure.ai (India).[10] Detailed overviews of each AI algorithm can be found in appendix 3 (pp 4–6).

Each centre was equipped with three four-module GeneXpert systems and all individuals were asked to submit a fresh spot sputum sample for testing with the Xpert MTB/RIF (Xpert) assay. An average of 12 Xpert tests were done daily at each centre. Xpert was repeated if the initial test failed (invalid, error, or no result). The final Xpert results were used as the bacteriological evidence and reference standard. All data collected were entered in a customised OpenMRS database and all chest x-ray images were anonymised using a pydicom module[22] in Python (script; appendix 3 pp 8–13).

The five AI algorithms scored the anonymised images retrospectively, independently, and blinded to all information, including age and sex. Three anonymised chest x-ray images were checked by the AI developers for image quality. No previous validation was done at the study site. We used the CAD4TB cloud version and installed the other four AI algorithms on the Stop TB Partnership server to analyse the anonymised chest x-ray files.

The AI developers of the five algorithms included in this study had no role in the study design, data collection, analysis plan, or writing of the study. The developers only had access to the chest x-ray images and did not receive any information on the patients' demographic, symptom, medical, or testing data.

Data analysis

We first compared the grouped performance of the group of three radiologists with the five AI algorithms to detect tuberculosis-suggestive abnormalities in individuals testing positive for *Mycobacterium tuberculosis* by Xpert. By dichotomising the categories used by the radiologists, we created three binary human reading classifications (A–C) varying the radiologist categories that were considered to be abnormal chest x-rays (appendix 3 p 14). To compare with the continuous output from AI, we calculated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the radiologists' three binary classifications and the threshold abnormality score each AI algorithm needed to match the sensitivity value for each one.[23] We then compared the difference in specificity, PPV, and NPV between human readings and those of the five AI algorithms using the McNemar test for paired proportions. We also compared each algorithm against WHO's Target Product Profile (TPP) for a triage tool of at least 90% sensitivity and at least 70% specificity by altering the threshold abnormality score to match each target value in turn and recording the performance.[8]

AUROCs were compared for each of the five algorithms using R's pROC package, using DeLong methods, for dependent AUROCs.[24] Since ROC plots can mislead on the reliability of algorithm performance owing to an intuitive but incorrect interpretation of the specificity in

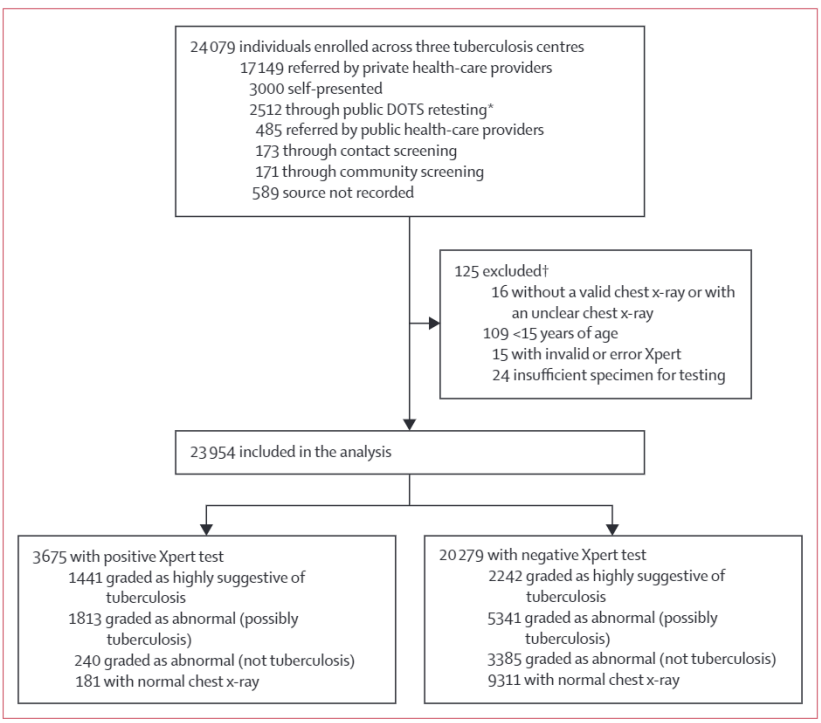

Figure 1: **Study flow diagram**

DOTS=directly observed treatment short course. Xpert=Xpert MTB/RIF. *DOTS centres are specialised facilities for the diagnosis and treatment of patients with tuberculosis. DOTS retesting refers to individuals referred to the study centres from a DOTS centre after a negative smear. †Individuals could be excluded for more than one reason.

	Overall (n=23 954)	Xpert results			Tuberculosis history		
		Positive (n=3675)	Negative (n=20 279)	p value	Previously treated (n=3586)	New case (n=20 341)	p value
Age, years	42·0 (30·0–57·0)	37·0 (27·0–53·0)	43·0 (31·0–58·0)	<0·0001	44·0 (31·0–58·0)	42·0 (30·0–57·0)	<0·0001
Age group				<0·0001			<0·0001
Young (15 to <25 years)	2666 (11·1%)	664 (18·1%)	2002 (9·9%)	..	309 (8·6%)	2355 (11·6%)	..
Middle aged (25 to <60 years)	16 056 (67·0%)	2378 (64·7%)	13 678 (67·4%)	..	2437 (68·0%)	13 606 (66·9%)	..
Older (≥60 years)	5232 (21·8%)	633 (17·2%)	4599 (22·7%)	..	840 (23·4%)	4380 (21·5%)	..
Sex				<0·0001			0·541
Female	7876 (32·9%)	1061 (28·9%)	6815 (33·6%)	..	1163 (32·4%)	6706 (33·0%)	..
Male	16 078 (67·1%)	2614 (71·1%)	13 464 (66·4%)	..	2423 (67·6%)	13 635 (67·0%)	..
Previous tuberculosis medication	3586 (15·0%)	608 (16·5%)	2978 (14·7%)	0·0041	3586 (100·0%)	0	..
Symptoms							
Cough	21 494 (89·7%)	3416 (93·0%)	18 078 (89·1%)	<0·0001	3176 (88·6%)	18 318 (90·1%)	0·0090
Fever	19 041 (79·5%)	3200 (87·1%)	15 841 (78·1%)	<0·0001	2931 (81·7%)	16 110 (79·2%)	0·0006
Shortness of breath	13 011 (54·3%)	2062 (56·1%)	10 949 (54·0%)	0·022	1888 (52·6%)	11 123 (54·7%)	0·024
Weight loss	15 035 (62·8%)	2774 (75·5%)	12 261 (60·5%)	<0·0001	2475 (69·0%)	12 560 (61·7%)	<0·0001
Haemoptysis	3103 (13·0%)	471 (12·8%)	2632 (13·0%)	0·807	511 (14·2%)	2592 (12·7%)	0·014
Any symptoms	23 582 (98·4%)	3644 (99·2%)	19 938 (98·3%)	0·0002	3499 (97·6%)	20 056 (98·6%)	<0·0001
Patient source				<0·0001			<0·0001
Community screening	170 (0·7%)	16 (0·4%)	154 (0·8%)	..	20 (0·6%)	150 (0·7%)	..
Contact screening	172 (0·7%)	9 (0·3%)	163 (0·8%)	..	11 (0·3%)	161 (0·8%)	..
Private referral	17 056 (71·2%)	2822 (78·7%)	14 234 (70·2%)	..	2515 (70·1%)	14 521 (71·4%)	..
Public DOTS retesting	2496 (10·7%)	436 (12·2%)	2060 (10·2%)	..	597 (16·6%)	1897 (9·3%)	..
Public referral	485 (2·1%)	100 (2·8%)	385 (1·9%)	..	47 (1·3%)	438 (2·2%)	..
Walk-in	2992 (12·8%)	204 (5·7%)	2788 (13·7%)	..	294 (8·2%)	2694 (13·2%)	..
Bacteriologically positive (Xpert)	3675 (15·3%)	3675 (100·0%)	0	..	608 (17·0%)	3063 (15·1%)	..
Mycobacterium tuberculosis burden				..			0·0004
Very low	634/3675 (17·3%)	634 (17·3%)	..	..	127/608 (20·9%)	507/3063 (16·6%)	..
Low	1093/3675 (29·7%)	1093 (29·7%)	..	..	192/608 (31·6%)	903/3063 (29·5%)	..
Medium	1299/3675 (35·3%)	1299 (35·3%)	..	..	187/608 (30·8%)	1111/3063 (36·3%)	..
High	648/3675 (17·6%)	648 (17·6%)	..	..	103/608 (16·9%)	546/3063 (17·8%)	..
Rifampicin resistance				..			<0·0001
Detected	181/3675 (4·9%)	181 (4·9%)	..	..	87/608 (14·3%)	94/3063 (3·1%)	
Not detected	3475/3675 (94·6%)	3475 (94·7%)	..	..	520/608 (85·5%)	2951/3063 (96·3%)	
Indeterminate	14/3675 (0·4%)	14 (0·4%)	..	..	1/608 (0·2%)	13/3063 (0·4%)	
Radiologist grading				<0·0001			<0·0001
Abnormal: highly suggestive	3683 (15·4%)	1441 (39·2%)	2242 (11·1%)	..	969 (27·0%)	2712 (13·3%)	..
Abnormal: possibly tuberculosis	7154 (29·9%)	1813 (49·3%)	5341 (26·3%)	..	1467 (40·9%)	5679 (27·9%)	..
Abnormal: not tuberculosis	3625 (15·1%)	240 (6·5%)	3385 (16·7%)	..	396 (11·0%)	3224 (15·8%)	..
Normal	9492 (39·6%)	181 (4·9%)	9311 (45·9%)	..	754 (21·0%)	8726 (42·9%)	..
AI abnormality scores							
CAD4TB*	0·17 (0·02–0·80)	0·97 (0·88–0·99)	0·09 (0·02–0·55)	<0·0001	0·61 (0·17–0·91)	0·12 (0·02–0·74)	<0·0001
qXR	0·24 (0·03–0·78)	0·89 (0·82–0·93)	0·11 (0·02–0·61)	<0·0001	0·68 (0·25–0·85)	0·15 (0·03–0·75)	<0·0001
Lunit INSIGHT CXR	0·29 (0·02–0·86)	0·95 (0·88–0·97)	0·10 (0·02–0·76)	<0·0001	0·81 (0·31–0·91)	0·15 (0·02–0·84)	<0·0001
JF CXR-1	0·85 (0·08–1·00)	1·00 (1·00–1·00)	0·58 (0·05–0·99)	<0·0001	1·00 (0·88–1·00)	0·70 (0·06–1·00)	<0·0001
InferRead DR	0·28 (0·14–0·65)	0·75 (0·59–0·83)	0·22 (0·13–0·54)	<0·0001	0·60 (0·31–0·75)	0·24 (0·13–0·61)	<0·0001

Data are n (%), n/N (%), or median (IQR), unless specified otherwise. Breakdown by radiologist grading is given in table 2. AI=artificial intelligence. DOTS=directly observed treatment short course. Xpert=Xpert MTB/RIF. *CAD4TB threshold abnormality scores have been standardised to aid comparison.

Table 1: **Characteristics of the 23 954 individuals included in this study, by Xpert results and tuberculosis history**

imbalanced datasets (ie, with low disease prevalence),[25] we also calculated the area under the precision–recall curve (PRC) for each algorithm (see appendix 3 p 15 for ROC and PRC methodology). Both ROC curves and PRCs were generated for each algorithm over a continuous range of threshold abnormality values.

We also assessed the distribution of abnormality scores disaggregated by Xpert results and patient history of tuberculosis. Finally, since the same threshold abnormality scores might provide different results in different populations, we evaluated the performance of the AI algorithms disaggregated by age, history of

	Overall (n=23 954)	Radiologist grading				
		Abnormal: highly suggestive of tuberculosis (n=3683)	Abnormal: possibly tuberculosis (n=7154)	Abnormal: not tuberculosis (n=3625)	Normal (n=9492)	p value
Age, years	42·0 (30·0–57·0)	45·0 (31·0–60·0)	47·0 (32·0–62·0)	54·0 (39·0–65·0)	35·0 (28·0–48·0)	<0·0001
Age group	..	..	..	..	..	<0·0001
Young (15 to <25 years)	2666 (11·1%)	377 (10·2%)	708 (9·9%)	182 (5·0%)	1399 (14·7%)	..
Middle aged (25 to <60 years)	16 056 (67·0%)	2363 (64·2%)	4488 (62·7%)	2048 (56·5%)	7157 (75·4%)	..
Older (≥60 years)	5232 (21·8%)	943 (25·6%)	1958 (27·4%)	1395 (38·5%)	936 (9·9%)	..
Sex	..	..	..	..	..	<0·0001
Female	7876 (32·9%)	1057 (28·7%)	2196 (30·7%)	1337 (36·9%)	3286 (34·6%)	..
Male	16 078 (67·1%)	2626 (71·3%)	4958 (69·3%)	2288 (63·1%)	6206 (65·4%)	..
Previous tuberculosis medication	3586 (15·0%)	969 (26·3%)	1467 (20·5%)	396 (10·9%)	754 (7·9%)	<0·0001
Symptoms						
Cough	21 494 (89·7%)	3376 (91·7%)	6443 (90·1%)	3331 (91·9%)	8344 (87·9%)	<0·0001
Fever	19 041 (79·5%)	3198 (86·8%)	5951 (83·2%)	2825 (77·9%)	7067 (74·5%)	<0·0001
Shortness of breath	13 011 (54·3%)	2287 (62·1%)	4064 (56·8%)	1953 (53·9%)	4707 (49·6%)	<0·0001
Weight loss	15 035 (62·8%)	2871 (78·0%)	4964 (69·4%)	2201 (60·7%)	4999 (52·7%)	<0·0001
Haemoptysis	3103 (13·0%)	597 (16·2%)	892 (12·5%)	452 (12·5%)	1162 (12·2%)	<0·0001
Any symptoms	23 582 (98·4%)	3652 (99·2%)	7039 (98·4%)	3581 (98·8%)	9310 (98·1%)	<0·0001
Patient source	..	..	..	..	..	<0·0001
Community screening	170 (0·7%)	28 (0·8%)	30 (0·4%)	15 (0·4%)	97 (1·0%)	..
Contact screening	172 (0·7%)	12 (0·3%)	31 (0·4%)	20 (0·6%)	109 (1·1%)	..
Private referral	17 056 (71·2 %)	2796 (75·9%)	5565 (78·8%)	2647 (73·0%)	6048 (63·7%)	..
Public DOTS retesting	2496 (10·7%)	510 (13·8%)	685 (9·6%)	282 (7·8%)	1019 (10·7%)	..
Public referral	485 (2·1%)	39 (1·1%)	212 (3·0%)	94 (2·6%)	140 (1·5%)	..
Walk-in	2992 (12·8%)	207 (5·6%)	492 (6·9%)	409 (11·3%)	1884 (19·8%)	..
Bacteriologically positive (Xpert)	3675 (15·3%)	1441 (39·1%)	1813 (25·3%)	240 (6·6%)	181 (1·9%)	<0·0001
Mycobacterium tuberculosis burden	..	..	..	..	..	<0·0001
Very low	634/3675 (17·3%)	192/1441 (13·3%)	304/1813 (16·8%)	60/240 (25·0%)	79/181 (43·6%)	..
Low	1093/3675 (29·7%)	391/1441 (27·1%)	570/1813 (31·4%)	82/240 (34·2%)	52/181 (28·7%)	..
Medium	1299/3675 (35·3%)	547/1441 (38·0%)	643/1813 (35·4%)	69/240 (28·8%)	41/181 (22·7%)	..
High	648/3675 (17·6%)	312/1441 (21·7%)	301/1813 (16·6%)	28/240 (11·7%)	9/181 (5%)	..
Rifampicin resistance	..	..	..	..	..	<0·0001
Detected	181/3675 (4·9%)	91/1441 (6·3%)	79/1813 (4·4%)	8/240 (3·3%)	3/181 (1·7%)	..
Not detected	3475/3675 (94·6%)	1347/1441 (93·5%)	1724/1813 (95·1%)	230/240 (95·8%)	174/181 (96·1%)	..
Indeterminate	14/3675 (0·4%)	1/1441 (0·1%)	8/1813 (0·4%)	2/240 (0·8%)	3/181 (1·7%)	..
Radiologist grading						
Abnormal: highly suggestive	3683 (15·4%)	..	..	..	..	..
Abnormal: possibly tuberculosis	7154 (29·9%)	..	..	..	..	..
Abnormal: not tuberculosis	3625 (15·1%)	..	..	..	..	..
Normal	9492 (39·6%)	..	..	..	..	..

(Table 2 continues on next page)

	Overall (n=23 954)	Radiologist grading				
		Abnormal: highly suggestive of tuberculosis (n=3683)	Abnormal: possibly tuberculosis (n=7154)	Abnormal: not tuberculosis (n=3625)	Normal (n=9492)	p value
(Continued from previous page)						
AI abnormality scores						
CAD4TB*	0·17 (0·02–0·80)	0·91 (0·65–0·98)	0·72 (0·34–0·95)	0·11 (0·03–0·45)	0·02 (0·01–0·06)	<0·0001
qXR	0·24 (0·03–0·78)	0·85 (0·74–0·91)	0·72 (0·41–0·86)	0·16 (0·04–0·51)	0·03 (0·02–0·07)	<0·0001
Lunit INSIGHT CXR	0·29 (0·02–0·86)	0·91 (0·83–0·96)	0·82 (0·51–0·93)	0·19 (0·03–0·64)	0·02 (0·01–0·05)	<0·0001
JF CXR-1	0·85 (0·08–1·00)	1·00 (1·00–1·00)	1·00 (0·96–1·00)	0·72 (0·21–0·98)	0·06 (0·02–0·32)	<0·0001
InferRead DR	0·28 (0·14–0·65)	0·72 (0·59–0·81)	0·59 (0·35–0·75)	0·23 (0·15–0·43)	0·13 (0·10–0·20)	<0·0001

Data are n (%), n/N (%), or median (IQR), unless specified otherwise. AI=artificial intelligence. DOTS=directly observed treatment short course. Xpert=Xpert MTB/RIF.
*CAD4TB threshold abnormality scores have been standardised to aid comparison.

Table 2: **Characteristics of the 23 954 individuals included in this study, by radiologist grading**

tuberculosis, sex, and patient sources using the AUROC and area under the PRC (PRAUC).

Evaluation framework

We used an evaluation framework that analyses performance beyond the AUROC alone, which is the standard approach of AI evaluations, to inform threshold selection by factoring in cost-effectivness (ie, reducing test expenditure) and the ability to triage. Algorithms were evaluated in a hypothetical triage process whereby the AI score output would be used to triage all individuals in the study population for follow-on Xpert diagnosis based on a predefined threshold abnormality score. We calculated the proportion of subsequent Xpert assays saved (with 0% representing the Xpert testing-for-all scenario) as a proxy for a product's cost-effectiveness. Likewise, the number of people needed to test (NNT) to find one bacteriologically positive individual was used as a proxy for a product's ability to triage. We plotted the sensitivity against the proportion of Xpert tests avoided to show the trade-off between finding as many bacteriologically positive patients as possible and the cost savings of each AI algorithm. We produced visualisations of sensitivity, proportion of Xpert tests avoided, and NNT over a continuous range of threshold abnormality scores in an evaluation framework to facilitate our understanding of threshold selection.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Between May 15, 2014, and Oct 4, 2016, 24 079 individuals visited the three study centres and were enrolled in this study. 109 were younger than 15 years of age. Xpert tests needed to be repeated in 830 participants: 15 tests remained invalid or showed an error after the second Xpert and 24 samples did not have enough specimen for the second Xpert. 16 individuals did not have a valid or clear x-ray. After excluding these 55 individuals, 23 954 (98·1%) individuals were included in the analysis (figure 1). The median age of participants was 42·0 years (30·0–57·0), a third were female, and almost all reported at least one tuberculosis-related symptom (tables 1, 2). Reported symptoms included cough (89·7%), fever (79·5%), shortness of breath (54·3%), weight loss (62·8%), and haemoptysis (13·0%). 3586 (15·0%) participants had previously received treatment for tuberculosis. More than three quarters of participants were referred by public or private health-care providers (tables 1, 2). The prevalence of bacteriologically positive tuberculosis confirmed by Xpert was 15·3% overall (n=3675), and 181 (4·9%) of these cases were resistant to rifampicin. Rifampicin resistance was observed in 87 (14·3%) of 608 Xpert-positive patients with a history of tuberculosis and 94 (3·1%) of 3063 without a history of tuberculosis. The radiologists graded 3683 (15·4%) radiographs as being highly suggestive of tuberculosis, 7154 (29·9%) as possibly tuberculosis, 3625 (15·1%) as abnormal but not tuberculosis, and 9492 (39·6%) as normal (table 2).

When the sensitivity of the AI algorithms was matched to that of the radiologists, all AI algorithms had significantly better specificity across the three binary classifications compared to the radiologists (table 3). Classification A (with only chest x-rays highly suggestive of tuberculosis being classified as radiologically positive) had a sensitivity of 38·9% (95% CI 37·3–40·5) and the highest specificity of the three classifications, at 88·9% (88·5–89·4); the AI algorithms' improvement in specificity at the same sensitivity level ranged from

	Threshold abnormality score*	Sensitivity*	Specificity	PPV	NPV	Absolute difference between AI and radiologists reading†		
						Specificity	PPV	NPV
Binary classification A								
Radiologists	..	38·9% (37·3 to 40·5)	88·9% (88·5 to 89·4)	39·1% (37·5 to 40·7)	89·0% (88·5 to 89·4)	..	..	..
AI algorithm								
CAD4TB‡	0·98	..	97·8% (97·6 to 98·0)	76·2% (74·2 to 78·1)	89·8% (89·4 to 90·2)	8·9% (8·4 to 9·4)	37·1% (36·2 to 38·0)	0·8% (0·2 to 1·4)
InferRead DR	0·79	..	94·2% (93·8 to 94·5)	54·9% (52·9 to 56·8)	89·4% (89·0 to 89·8)	5·2% (4·7 to 5·8)	15·7% (14·8 to 16·7)	0·4% (–0·2 to 1·0)
JF CXR-1	1·00	..	93·5% (93·1 to 93·8)	54·2% (52·4 to 56·0)	89·9% (89·5 to 90·3)	4·6% (4·0 to 5·1)	15·1% (14·1 to 16·0)	1·0% (0·4 to 1·6)
Lunit INSIGHT CXR	0·96	..	98·0% (97·8 to 98·1)	75·5% (73·3 to 77·5)	89·1% (88·7 to 89·5)	9·0% (8·5 to 9·5)	36·3% (35·4 to 37·2)	0·2% (–0·5 to 0·8)
qXR	0·91	..	97·9% (97·7 to 98·1)	75·9% (73·8 to 77·8)	89·5% (89·1 to 89·9)	8·9% (8·5 to 9·4)	36·8% (35·9 to 37·7)	0·5% (–0·1 to 1·1)
Binary classification B								
Radiologists	..	88·5% (87·4 to 89·5)	62·5% (61·8 to 63·1)	30·0% (29·2 to 30·9)	96·8% (96·5 to 97·1)	..	..	..
AI algorithm								
CAD4TB‡	0·57	..	75·8% (75·2 to 76·4)	40·0% (39·0 to 41·1)	97·3% (97·0 to 97·5)	13·4% (12·5 to 14·3)	10·0% (9·1 to 10·9)	0·51% (0·2 to 0·8)
InferRead DR	0·37	..	64·5% (63·8 to 65·1)	31·2% (30·3 to 32·1)	96·8% (96·5 to 97·1)	2·0% (1·1 to 3·0)	1·2% (0·3 to 2·1)	0·1% (–0·3 to 0·4)
JF CXR-1	0·95	..	64·1% (63·4 to 64·7)	31·0% (30·1 to 31·9)	96·8% (96·5 to 97·1)	1·6% (0·7 to 2·6)	1·0% (0·1 to 1·9)	0·0% (–0·3 to 0·4)
Lunit INSIGHT CXR	0·66	..	70·3% (69·7 to 71·0)	35·3% (34·3 to 36·3)	97·1% (96·8 to 97·4)	7·9% (7·0 to 8·8)	5·3% (4·4 to 6·2)	0·3% (0·0 to 0·6)
qXR	0·64	..	76·7% (76·1 to 77·2)	40·9% (39·8 to 41·9)	97·4% (97·1 to 97·6)	14·2% (13·3 to 15·1)	10·8% (9·9 to 11·8)	0·6% (0·2 to 0·9)
Binary classification C								
Radiologists	..	95·0% (94·3 to 95·7)	45·7% (45·0 to 46·4)	24·2% (23·5 to 24·9)	98·1% (97·8 to 98·4)	..	..	..
AI algorithm								
CAD4TB‡	0·18	..	58·5% (57·8 to 59·1)	29·5% (28·7 to 30·3)	98·5% (98·2 to 98·7)	12·8% (11·9 to 13·8)	5·3% (4·5 to 6·2)	0·4% (0·1 to 0·6)
InferRead DR	0·20	..	47·5% (46·8 to 48·2)	25·0% (24·2 to 25·7)	98·1% (97·9 to 98·4)	1·8% (0·8 to 2·8)	0·8% (0·0 to 1·7)	0·1% (–0·2 to 0·3)
JF CXR-1	0·53	..	49·0% (48·3 to 49·7)	25·5% (24·8 to 26·2)	98·2% (97·9 to 98·5)	3·3% (2·3 to 4·3)	1·3% (0·5 to 2·2)	0·1% (–0·2 to 0·4)
Lunit INSIGHT CXR	0·07	..	47·8% (47·1 to 48·5)	25·6% (24·8 to 26·3)	98·2% (97·9 to 98·5)	2·2% (1·2 to 3·1)	1·4% (0·5 to 2·2)	0·1% (–0·2 to 0·4)
qXR	0·35	..	63·5% (62·9 to 64·2)	32·2% (31·3 to 33·1)	98·6% (98·4 to 98·8)	17·9% (16·9 to 18·8)	8·0% (7·2 to 8·9)	0·5% (0·2 to 0·7)

Data in parentheses are 95% CIs. Binary classifications used can be found in appendix 3 (p 14). AI=artificial intelligence. NPV=negative predictive value. PPV=positive predictive value. *Threshold abnormality scores were chosen to match the sensitivity across all AI algorithms to that obtained by the three radiologists. †Data are percentage point differences. A positive difference means that the specificity, PPV, or NPV of the AI algorithm is higher than that of the radiologists, when matching sensitivity. A negative difference means that the specificity, PPV, or NPV of the AI algorithm is lower than that of the radiologists, when matching sensitivity. For example, the difference in specificity=the specificity of an AI algorithm – the specificity of the corresponding radiological binary classification. ‡CAD4TB threshold abnormality scores have been standardised to aid comparison.

Table 3: **Comparison of sensitivity and specificity between radiologists' reading and the predictions of the AI algorithms by human binary classifications**

Lunit INSIGHT CXR's 9·0 percentage points (95% CI 8·5–9·5) to JF CXR-1's 4·6 percentage points (4·0–5·1). Additionally, all AI algorithms significantly improved on the human readers' PPV, except for InferRead DR when compared with radiological classification C (with only normal chest x-ray being classified as radiologically negative; table 3). However, for all three radiological binary classifications, the difference in NPVs between most AI algorithms and human readers were not significant (table 3).

At 90% sensitivity, algorithm specificities were highest for qXR and lowest for JF CXR-1 (table 4).

	Threshold score	Sensitivity (95% CI)	Specificity (95% CI)
Sensitivity fixed at 90%			
CAD4TB*	0·50	90·0% (89·0–91·0)	72·9% (72·3–73·5)
InferRead DR	0·34	90·3% (89·3–91·3)	62·1% (61·4–62·7)
JF CXR-1	0·92	90·4% (89·4–91·3)	61·1% (60·4–61·8)
Lunit INSIGHT CXR	0·60	90·1% (89·0–91·0)	67·2% (66·6–67·9)
qXR	0·60	90·2% (89·2–91·1)	74·3% (73·3–74·9)
Specificity fixed at 70%			
CAD4TB*	0·44	91·5% (90·5–92·4)	70·0% (69·4–70·6)
InferRead DR	0·47	84·0% (82·8–85·2)	70·6% (69·9–71·2)†
JF CXR-1	0·98	85·0% (83·8–86·2)	68·8% (68·2–69·5)†
Lunit INSIGHT CXR	0·67	88·8% (87·7–89·8)	70·1% (69·4–70·7)
qXR	0·51	92·6% (91·7–93·4)	70·3% (69·6–70·9)

AI=artificial intelligence. *CAD4TB threshold abnormality scores have been standardised to aid comparison. †Marks the closest available match to a specificity value of 70%.

Table 4: **Comparison of AI algorithms against WHO's Target Product Profile when matching either sensitivity or specificity**

At 70% specificity, algorithm sensitivities were highest for qXR and lowest for InferRead DR. Only qXR and CAD4TB met the TPP of at least 90% sensitivity and at least 70% specificity, with Lunit INSIGHT CXR coming close (table 4).

The trade-offs between sensitivity and specificity of the five AI algorithms can be visualised in the ROC curves and PRCs (figure 2). The AUROCs were 90·81% (95% CI 90·33–91·29) for qXR, 90·34% (89·81–90·87) for CAD4TB, 88·61% (88·03–89·20) for Lunit INSIGHT CXR, 84·90% (84·27–85·54) for InferRead DR, and 84·89% (84·26–85·53) for JF CXR-1. The AUROCs of CAD4TB and qXR were significantly greater than the others (p values $<2·2\times10^{-16}$; appendix 3 p 15). However, no significant difference in the AUROCs was observed between JF CXR-1 and InferRead (p=0·97; appendix 3 p 15). Above the 90% sensitivity mark, the ROC curves across AI algorithms did not differ significantly. By contrast, the PRCs and corresponding PRAUC scores (CAD4TB: 66·95%; qXR: 66·46%; Lunit INSIGHT CXR: 62·20%; JF CXR-1: 50·86%; and InferRead DR: 49·56%) show a clear difference between the five algorithms, with certain AI algorithms having lower precision values for some given recall (ie, sensitivity)

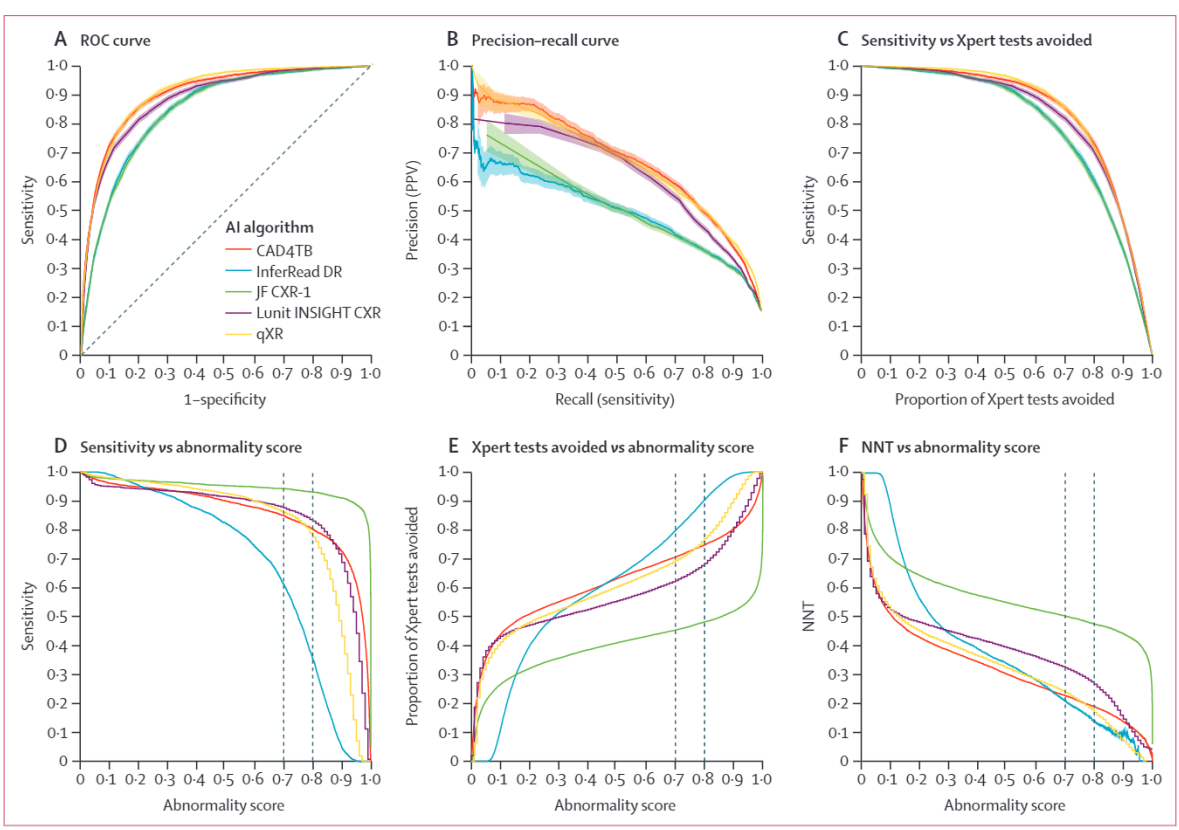

Figure 2: **Performance metrics between AI algorithms**

Measures of performance include the ROC curves (A), precision–recall curves (B), and trade-off between sensitivity and proportion of subsequent Xpert tests avoided (C) for the five AI algorithms, as well as sensitivity (D), proportion of subsequent Xpert tests avoided (E), and NNT (F) over a continuous range of threshold abnormality scores. Performance is based on data from 23 954 individuals. AI=artificial intelligence. NNT=number needed to test. PPV=positive predictive value. ROC=receiver operating characteristic. Xpert=Xpert MTB/RIF.

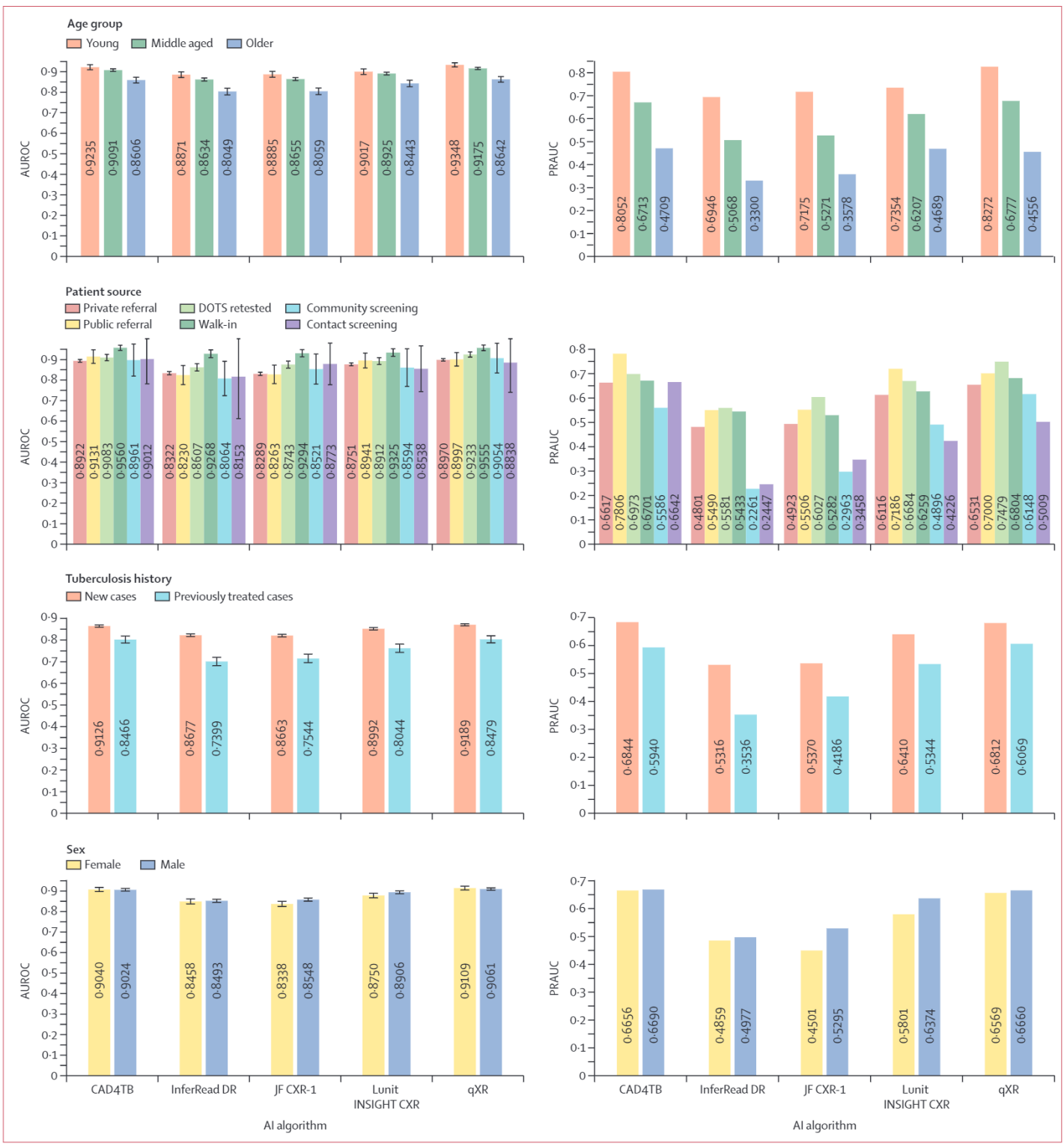

Figure 3: **AUROC and PRAUC by subgroups**

Measures are disaggregated by age (young [15 to <25 years], middle aged [25 to <60 years], and older [≥60 years]); patient source; tuberculosis history; and sex. AI=artificial intelligence. AUROC=area under the receiver operating characteristic curve. DOTS=directly observed treatment short course. PRAUC=area under the precision–recall curve.

values (figure 2). Example report outputs from AI algorithms are provided in appendix 3 (pp 16–20).

All five AI algorithms were found to reduce the number of Xpert tests required by 50% while maintaining a sensitivity above 90% (figure 2C). However, as more diagnostic tests are avoided (especially >60%), significant differences in sensitivity become apparent between some of the algorithms. When Xpert testing is reduced by two thirds, the sensitivity across all algorithms ranged from 80% to 88%.

Although the ROC curves and PRCs showed similar performance between InferRead DR and JF CXR-1, the evaluation framework showing the dynamics of sensitivity, proportion of Xpert tests avoided, and NNT with varying threshold abnormality scores revealed significant differences (figure 2D–F). For most of the decision thresholds (above approximately 0·15), JF CXR-1 had higher sensitivity, but avoided fewer Xpert tests and required a higher NNT than did InferRead DR. For instance, at a cutoff threshold of 0·8, JF CXR-1 has a sensitivity of 93·0% (95% CI 92·1–93·8), avoids 48·7% of Xpert tests, and has an NNT of 3·6 (3·5–3·7). At the same threshold, InferRead DR has a sensitivity of 35·4% (33·9–37·0), avoids 90·5% of Xpert tests, and has an NNT of 1·8 (1·7–1·8).

The performance of a single AI algorithm across the evaluation framework can be used to inform threshold selection. For most threshold abnormality scores (ie, excluding very low or high scores), the sensitivity of JF CXR-1 remained above 90%, the percentage of Xpert tests avoided was between 25% and 60%, and the NNT was between five and three (figure 2). The sensitivity of CAD4TB, Lunit INSIGHT CXR, and qXR remained higher than approximately 80% for most of the threshold scores (≤0·8), before quickly decreasing (figure 2). The threshold selection clearly depends on the algorithm in question and the context in which it is being used. For example, the threshold required to achieve at least 90% sensitivity must be below 0·34 for InferRead DR, below 0·50 for CAD4TB, below 0·60 for qXR and Lunit INSIGHT CXR, and below 0·93 for JF CXR-1 (figure 2D).

The distributions of the abnormality scores of the five AI algorithms disaggregated by Xpert outcomes and history of tuberculosis vary considerably, possibly indicating different underlying neural networks and the effect that changing the threshold abnormality score can have for different algorithms (appendix 3 p 21). Density plots for Lunit INSIGHT CXR, CAD4TB, qXR's, and InferRead DR showed a good dichotomisation pattern (between bacteriologically positive and negative). Although almost all bacteriologically positive participants received high abnormality scores (0·95–1·00) from JF CXR-1, so did many bacteriologically negative individuals. No abnormality score distributions for bacteriologically negative participants with a history of tuberculosis (the dark red bars) are left skewed for any AI algorithm (appendix 3 p 21).

All five AI algorithms showed significant variation in performance with respect to age, performing worse in the older age group (>60 years) than in both the young group (p values ranging from $1·8\times10^{-16}$ [qXR] to $9·6\times10^{-8}$ [Lunit INSIGHT CXR]) and middle-aged group (p values ranging from $6·7\times10^{-13}$ [qXR] to $6·7\times10^{-8}$ [Lunit INSIGHT CXR]; figure 3; appendix 3 pp 22–23). InferRead DR, CAD4TB, JF CXR-1, and qXR also performed significantly worse in the middle-aged group compared with the younger group, although no significance was observed for Lunit INSIGHT CXR. All five AI algorithms performed significantly worse among people with a history of tuberculosis (p values from $1·6\times10^{-30}$ [InferRead DR] to $1·1\times10^{-12}$ [CAD4TB]; figure 3; appendix 3 p 22). No significant differences were observed between the sexes, except JF CXR-1 and Lunit INSIGHT CXR, which performed better in males than females (p=0·0045 [JF CXR-1] and p=0·020 [Lunit INSIGHT CXR; figure 3; appendix 3 p 22).

AI performance also varied with patient source. All algorithms performed significantly better among walk-ins than referrals from public and private facilities, and directly observed treatment short course (DOTS) retesting (p values ranging from $7·5\times10^{-25}$ [*vs* private referral, with JF CXR-1] to 0·017 [*vs* public referral, with CAD4TB]), except Lunit INSIGHT CXR which showed no difference with public sector referral (figure 3; appendix 3 p 23). qXR (p=0·0002), JF CXR-1 ($p=3·0\times10^{-6}$), and InferRead DR (p=0·0039) performed better among individuals from DOTS retesting than private referrals, and InferRead DR also performed significantly worse among individuals from community screening than walk-ins (p=0·0065; figure 3; appendix 3 pp 23–24). Performance among individuals from contact screening did not significantly differ from any of the other patient sources across all AI algorithms (appendix 3 pp 23–24).

Discussion

This is the largest independent study evaluating multiple AI algorithms as triage tests for tuberculosis with chest x-ray, and the first published evaluation of JF CXR-1, InferRead DR, and the latest version of CAD4TB (version 7) for detecting tuberculosis-suggested abnormalities. Our study shows that the predictions made by the five algorithms significantly outperformed experienced human readers in detecting tuberculosis-related abnormalities.[14]

The AUROCs indicate that all AI algorithms performed well, with qXR and CAD4TB being the two top performers. CAD4TB, Lunit INSIGHT CXR, and qXR's AUROCs in this study are slightly lower than those in previous independent evaluations.[17,26] Our subanalysis showed that the performance of computer-aided detection varied with demographic and clinical factors, as well as patient source, and therefore implied that variation exists in AI performance across different contexts and geographies. In our study population, rates of bacteriological positivity and rifampicin resistance are both higher than the national average because of the high proportion of

referrals and urban population in this study.[8] Such findings caution the generalisation to other populations, particularly when selecting threshold abnormality scores for different populations. The different training strategies and datasets used (appendix 3 pp 4–6) could explain these differences in performance by affecting the ability of AI algorithms to generalise learning to different populations. Furthermore, we found that the difference in performance between algorithms is better visualised in a PRC plot than in ROC curves, so this approach should be used in future analyses with imbalanced datasets to inform software selection.

It is important that implementers can make informed decisions when selecting the threshold abnormality score specific to their settings. To enable this, we used a new evaluation framework with important implementation-relevant indications, such as number of confirmation tests avoided and NNT, to infer the cost-effectiveness and the ability to triage. We observed that automated reading of chest x-ray by all five AI algorithms can keep sensitivity above 90% and at least halve the number of follow-on diagnostic tests required. For large case-finding programmes that might have insufficient on-site Xpert testing capacity for all individuals examined, there is a trade-off between the number of cases identified and the proportion of Xpert tests that can be avoided—for example, choosing to avoid using 70–80% of Xpert tests by allowing sensitivity to reduce to 70% could miss 30% of tuberculosis cases. Our evaluation framework also indicates the difference in algorithms that might have very similar performance if assessed using ROC curves and PRC alone.

The importance of using a more nuanced analytical framework to evaluate performance can be illustrated by imagining different case-finding situations. For a programme focused on capturing almost all people with tuberculosis (eg, ≥95%), 46% of people triaged by qXR would be recalled for confirmatory tests, followed by 49% by CAD4TB, 57% by JF CXR-1, 58% by Lunit INSIGHT CXR, and 59% by InferRead DR. If, instead, we imagine a large active case-finding programme using chest x-ray but with a much smaller budget and the need to reduce the numbers of follow-on Xpert tests by 75% while accepting compromised sensitivity, the strongest candidates would be qXR (sensitivity 80·6%, 95% CI 79·2–81·8) and CAD4TB (79·7%, 78·3–81·0), followed by Lunit INSIGHT CXR (76·6%, 75·1–77·9), InferRead DR (69·3%, 67·7–70·8), and JF CXR-1 (8·5%, 66·6–69·7). We recommend that this evaluation framework be included in future AI evaluations instead of only reporting on AUROCs.

Our density plots show that the underlying neural networks of the five AI algorithms might be constructed very differently and that no universal threshold abnormality score can be applicable to all algorithms. Moreover, the density plots of bacteriologically negative individuals with a history of tuberculosis indicates the algorithms' poor ability to differentiate between old scarring and active lesions could lead to excessive recall in this group. We hypothesise that the abnormalities on the chest due to age and previous tuberculosis influenced the classification of active tuberculosis. The overall performance of the five AI algorithms differed between age groups, patient sources, and history of tuberculosis, although the algorithms performed similarly for both sexes. Interestingly, accuracy (ie, AUROC) was lowest among those who presented themselves and were recruited through community-based case finding. This implies that the threshold abnormality scores probably need to be different depending on the population tested and with subpopulations. We further recommend that manufacturers include basic demographic and clinical data when training the AI algorithms to further improve their algorithms in future software iterations.

Our results document the performance of five algorithms at one point in time. New algorithms will probably emerge in the near future and updated software versions are launched almost annually.[10] Two algorithms in this evaluation had not been previously evaluated in peer-reviewed journals. Unlike traditional diagnostic tests, which take years to produce and update, the performance of AI improves incredibly fast. Future guidance from bodies such as WHO must prepare for this speed of change and independent evaluation libraries are required to help implementers to understand the performance of the latest offerings in the field.

Our study has several limitations. Due to logistic and budgetary constraints, we did not use culture as the reference standard, meaning that some people with Xpert-negative, culture-positive tuberculosis might have incorrectly been labelled as not having tuberculosis. We also did not have access to Xpert Ultra, which is more sensitive than Xpert, in Bangladesh during the study. Due to the small number of asymptomatic individuals (1·6% of the evaluation population), we did not stratify by symptoms and analyse by symptom subgroup. We did not do any HIV testing because Bangladesh has a low HIV prevalence.[4] However, algorithm performance among different subpopulations—especially those living with HIV, who often present with atypical radiological images—needs to be better documented.[27] Similarly, we excluded children from our study population, even though some (but not all) of the algorithms included are licensed for use in younger age groups (appendix 3 pp 4–6). These decisions limit the generalisability of our findings. In particular, further evaluation of computer-aided detection in children is necessary. Another limitation is that each chest x-ray was read by a single radiologist, rather than a panel of radiologists. However, the intended use of these AI algorithms is in resource-constrained settings with few or no radiologists and neither resources nor time permitted multiple readings of high numbers of images. Furthermore, human readers

were blinded to clinical and demographic information except for age and sex, although in the field, reading of chest x-rays could be informed by this data. Additionally, the analysis included only three human readers as a comparison point. We caution against extrapolating the study findings to rural areas as our study was done in metropolitan Dhaka, where experienced readers are often more available. Indeed, computer-aided detection software might perform even better compared with radiologists in rural and poorly resourced areas. Additionally, only one brand of x-ray machine was used in this study due to procurement constraints. Lastly, we did not conduct this study prospectively and did not collect implementation data such as programmatic costs, setup, services, or user experience.

In conclusion, our results show that all five AI algorithms outperformed experienced certified radiologists and could avoid follow-on Xpert testing and reduce the NNT while maintaining high sensitivity. ROC curves and PRCs are powerful tools for evaluation; however, additional metrics and analysis, including our new evaluation framework of sensitivity, confirmation tests saved, and NNT with varying threshold abnormality scores, will help implementers with threshold and software selection.

Contributors

The study was conceived by ZZQ, JC, and SB. Data collection was led by SB, KP, SA, MSS, and ASSA. Data cleaning and verification was done by ZZQ, SB, KP, and SA; and data analysis and interpretation by ZZQ, TN, and JC. ZZQ wrote the first draft of the manuscript. ZZQ, JC, and RB revised the manuscript. ZZQ and SB had access to all data. All authors contributed to and approved the final manuscript.

Declaration of interests

We declare no competing interests.

Data sharing

The anonymised datasets used in this study can be available upon reasonable request to the corresponding author. Chest x-ray images will not be provided as these are withheld by the corresponding author's organisation to reserve their use for product evaluations.

Acknowledgments

Funding support of this project came from Global Affairs Canada through the Stop TB Partnership's TB REACH initiative (grant number STBP/TBREACH/GSA/W5-24). ZZQ, JC, TN, and RB work for the Stop TB Partnership. The authors were allowed to use the AI algorithms free of charge by all five AI companies for research purposes, but the companies had no influence over the research question, nor any other aspect of the work carried out, and had no impact on the transparency of the Article. The authors thank Pauline Vandewalle and Toufiq Rahman (Stop TB Partnership, Geneva, Switzerland) for providing translations of the abstract.

References

1 Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics* 2017; **37**: 2113–31.

2 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; **2012**: 1097–105.

3 WHO. Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches. Geneva: World Health Organization, 2016.

4 WHO. Global tuberculosis report 2020. Geneva: World Health Organization, 2020.

5 Nathavitharana RR, Yoon C, Macpherson P, et al. Guidance for studies evaluating the accuracy of tuberculosis triage tests. *J Infect Dis* 2019; **220** (suppl 3): S116–25.

6 Rahman M, Flora MS, Husain MM, et al. National tuberculosis prevalence survey 2015–2016. Dhaka: Ministry of Health and Family Welfare, 2016.

7 Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press, 2016.

8 WHO. High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28–29 April 2014. Geneva: World Health Organization, 2014.

9 Qin ZZ, Naheyan T, Ruhwald M, et al. A new resource on artificial intelligence powered computer automated detection software products for tuberculosis programmes and implementers. *Tuberculosis (Edinb)* 2021; **127**: 102049.

10 The Stop TB Partnership. FIND. Resource center on computer-aided detection products for the diagnosis of tuberculosis. 2020. https://www.ai4hlth.org/ (accessed Sept 10, 2020).

11 WHO. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. Geneva: World Health Organization, 2021.

12 Murphy K, Habib SS, Zaidi SMA, et al. Computer aided detection of tuberculosis on chest radiographs: an evaluation of the CAD4TB v6 system. *Sci Rep* 2020; **10**: 5492.

13 Zaidi SMA, Habib SS, Van Ginneken B, 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**: 12339.

14 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**: 1226–30.

15 Larson DB, Harvey H, Rubin DL, Irani N, Justin RT, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. *J Am Coll Radiol* 2021; **18**: 413–24.

16 Nash M, Kadavigere R, Andrade J, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. *Sci Rep* 2020; **10**: 210.

17 Qin ZZ, Sander MS, Rai B, 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**: 15000.

18 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997; **30**: 1145–59.

19 Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Croat Med J* 2003; **44**: 635–38.

20 Banu S, Haque F, Ahmed S, et al. Social Enterprise Model (SEM) for private sector tuberculosis screening and care in Bangladesh. *PLoS One* 2020; **15**: e0241437.

21 WHO. Tuberculosis prevalence surveys: a handbook. Geneva: World Health Organization, 2011.

22 Mason D. SU-E-T-33: pydicom: an open source DICOM library. *Med Phys* 2011; **38**: 3493.

23 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577**: 89–94.

24 Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**: 77.

25 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; **10**: e0118432.

26 Fehr J, Konigorski S, Olivier S, et al. Computer-aided interpretation of chest radiography to detect TB in rural South Africa (version 2). *medRxiv* 2020; published online Sept 9. https://doi.org/10.1101/2020.09.04.20188045 (preprint).

27 Padyana M, Bhat RV, Dinesha M, Nawaz A. HIV-tuberculosis: a study of chest x-ray patterns in relation to CD4 count. *N Am J Med Sci* 2012; **4**: 221–25.