Test Report

qXR

โดย Qure.Ai Technologies Inc.

รายงานผลการทดสอบ
โดยราชวิทยาลัยรังสีแพทย์แห่งประเทศไทย

ทดสอบใช้กับภาพรังสีทรวงอก ในกรณี

- คัดกรอง (screening) วัณโรคปอด
- อ่านผลซ้ำ (double reading) ให้กับรังสีแพทย์ เพื่อเพิ่มคุณภาพการวินิจฉัย
- เพิ่มความแม่นยำในการค้นหาพยาธิสภาพให้กับรังสีแพทย์
- ประมาณความยาก-ง่ายในการแปลผล
- จัดลำดับความเร่งด่วน (triage) ในการแปลผลให้แก่รังสีแพทย์

Report on the Test Performance of Artificial Intelligence for Tuberculosis Screening in Chest X-Ray Images of the Thai Population

Filer Name

Company	Qure.ai
Address	Level 6, Oberoi Commerz II International Business Park Oberoi Garden City, Yashodham, Goregaon, Mumbai, Maharashtra 400063, India
Contact	Bunty Kundnani and Ankolika Bhatia

Developer Company

Company	Qure.ai
Address	Level 6, Oberoi Commerz II International Business Park Oberoi Garden City, Yashodham, Goregaon, Mumbai, Maharashtra 400063, India
Country	India
Website	https://www.qure.ai

Software

Name	Qure.ai qXR
Version	Not provided by the developer
Description	Product specifications excerpted from https://grand-challenge.org/aiforradiology/ qXR detects abnormal chest X-rays, then identifies and localizes upto 29 common abnormalities. It also screens for tuberculosis. Product specifications — Information source: Vendor Last updated: Sept. 30, 2023

General	
Product name	qXR
Company	Qure.ai
Subspeciality	Chest
Modality	X-ray
Disease targeted	Tuberculosis, Covid-19, consolidation, fibrosis, blunted CP, pleural effusion, hilar enlargement, nasogastric and endotracheal tube detection, pneumothorax, pneumo peritoneum, rib fracture, nodule, lung opacities, cavity.
Key-features	Abnormality detection and localization, report generation, tuberculosis screening, worklist prioritization

	Suggested use	Before: flagging acute findings During: perception aid (prompting all abnormalities/results/heatmaps), report suggestion
	Data characteristics	
	Population	All chest X-rays
	Input	PA/ AP view chest X-rays
	Input format	DICOM
	Output	Image annotations, free text draft radiology reports
	Output format	DICOM
	Technology	
	Integration	Integration in standard reading environment (PACS), Integration RIS (Radiological Information System), Integration via AI marketplace or distribution platform, Stand-alone webbased
	Deployment	Locally on dedicated hardware, Locally virtualized (virtual machine, docker), Cloud-based
	Trigger for analysis	Automatically, right after the image acquisition, On demand, triggered by a user through e.g. a button click, image upload, etc.
	Processing time	10 - 60 seconds
	Certification	
	CE	Certified, Class IIb, MDR
	FDA	FDA 510(k) clearance only for facilitating confirmation of the position of the breathing tube, automated cardiothoracic ratio measurements, and triage of pneumothorax and pleural effusion.
	Market presence	
	On market since	05-2018
	Distribution channels	Nuance PIN, Incepto, Philips IntelliSpace, Sectra Amplifier Store, Blackford, GE Healthcare, Siemens

Dataset

Reference No.	1A2A–QureAI
Number of Images	806
Internal Validation	Consistent

Data Characteristics

The dataset consists of 806 randomly selected chest radiographic images from a pool of 1,500 images carefully curated from Songklanagarind Hospital in Songkhla Province, Chiangrai Pracharuk Hospital in Chiang Rai Province, Udon Thani Hospital in Udon Thani Province, Suttawet Hospital in Maha Sarakham Province, and the Tuberculosis Division of the Department of Disease Control, Ministry of Public Health. Each image was read by three B Readers. Our goal is to utilize high-quality datasets that are read by B Readers, who are trained and certified radiologists.

A B Reader is a qualified radiologist who is certified by the National Institute for Occupational Safety and Health (NIOSH) in the United States. B Readers are specifically trained to interpret and classify chest radiographs for the presence of pneumoconiosis, a group of lung diseases.

Characteristics of the radiographic images:

- Chest radiographic images of patients aged 15 years and above were included, taken with a computed radiography machine.
- No images from patients with a positive HIV Serology status.
- No images from patients with other opportunistic pulmonary infections or co-infections, such as Mycobacterium tuberculosis, Histoplasmosis, Cryptococcosis, Melioidosis, and Acinetobacter baumannii.

To assess the inter-rater reliability, the following metrics were employed:

- Pairwise Agreement: The average level of agreement among each pair of B readers.
- Intraclass Agreement (ICC): The average Pearson's correlation using ICC(2,3) when three B readers read the randomly selected radiographic images.
- Pairwise Cohen's Kappa and Fless' Kappa statistics for the analysis of agreement between assessors

Number of Findings

Table 1 presents the number of findings annotated by B Readers for chest X-ray images in Dataset 1A2A–QureAI, which consists of 806 images. Each image in the dataset was independently assessed by three randomly selected B Readers from a pool of six B Readers. $N_{\text{Individual Reader}}$ represents The number of findings that each individual B reader labelled, while $N_{\text{Concensus}}$ represents the number of findings where the majority of the B Readers agreed.

Table 1 Number of findings annotated annotated by B Readers in Dataset 1A2A–QureAI

Finding	$N_{Individual\ Reader}$	$N_{Concensus}$
Abnormalities	1,569	513
Small opacity	1,248	420
Primary nodular	925	323
Primary reticular	308	58
Secondary nodular	715	241
Secondary reticular	454	110
Large opacity	1,234	420
Mass/nodule	496	136
Cavity	875	296
Fibrosis	738	241
Calcification	298	58
Pleural effusion	327	109
Pleural thickening	555	179
Pneumothorax	14	4
Hilar adenopathy	315	72
Mediastinal adenopathy	95	17
Consistent with tuberculosis	1,264	414
Active Tuberculosis	1,216	406
Patchy infiltration	926	334
Cavity with surrounding consolidation	807	278
Unilateral hilar/paratracheal lymph node enlargement	147	30
Pleural effusion	165	49
Miliary nodules	310	76
Indeterminate tuberculosis	48	6
Reticulonodular infiltration	28	4
Destroyed lung or bronchiectasis	5	0
Inconsistent with tuberculosis	1,154	392

Inter-rater Reliability

Table 2 Inter-rater reliability measures for each finding in Dataset 1A2A–QureAI (806 images). Each finding was interpreted by three B Readers. The reliability was measured using statistical metrics such as Pairwise Agreement, ICC(2,3), Pairwise Cohen's kappa, and Fleiss' kappa.

Finding	Agreement	ICC	Cohen's	Fleiss'
Abnormalities	0.9222	0.9344	0.8294	0.8257
Small opacity	0.8594	0.8846	0.7185	0.7185
Primary nodular	0.8280	0.8399	0.6358	0.6358
Primary reticular	0.8065	0.3086	0.1292	0.1294
Secondary nodular	0.7072	0.5594	0.2969	0.2970
Secondary reticular	0.7436	0.3629	0.1594	0.1593
Large opacity	0.9041	0.9267	0.8081	0.8080
Mass/nodule	0.7750	0.5746	0.3100	0.3101
Cavity	0.8685	0.8831	0.7156	0.7152
Fibrosis	0.7643	0.7064	0.4445	0.4442
Calcification	0.8180	0.3605	0.1581	0.1580
Pleural effusion	0.9388	0.8945	0.7380	0.7383
Pleural thickening	0.8462	0.7961	0.5653	0.5650
Pneumothorax	0.9967	0.8816	0.7095	0.7126
Hilar adenopathy	0.8404	0.5581	0.2966	0.2955
Mediastinal adenopathy	0.9404	0.4479	0.2086	0.2111
Consistent with tuberculosis	0.9601	0.9720	0.9185	0.9204
Active Tuberculosis	0.9537	0.9671	0.9074	0.9074
Patchy infiltration	0.8296	0.8420	0.6395	0.6395
Cavity with surrounding consolidation	0.8503	0.8556	0.6634	0.6634
Unilateral hilar/paratracheal lymph node enlargement	0.9049	0.3761	0.1650	0.1670
Pleural effusion	0.9404	0.7733	0.5314	0.5317
Miliary nodules	0.8230	0.4413	0.2085	0.2082
Indeterminate tuberculosis	0.9686	0.4183	0.1969	0.1923
Reticulonodular infiltration	0.9801	0.3178	0.1642	0.1328
Destroyed lung or bronchiectasis	0.9959	-0.006	-0.0019	-0.0021
Inconsistent with tuberculosis	0.9603	0.9720	0.9204	0.9204

Table 3 Interpretation of ICC and Kappa Values according to Landis and Koch (1977)[1]

ICC/Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

[1] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. In Biometrics (Vol. 33, Issue 1, p. 159). JSTOR. https://doi.org/10.2307/2529310

Results

The inter-rater reliability is measured using Pairwise Agreement, which is the average similarity between each pair of B Readers and Qure.ai qXR, as well as Pairwise Cohen's Kappa, which is the average of Cohen's Kappa statistics between each pair of B Readers and Qure.ai qXR. This is done to compare the agreement between B Readers and Qure.ai qXR ("B" vs AI) and among B Readers themselves ("B" vs "B").

Table 4 Reliability Measures Within B Readers (“B” vs “B”) and Between the System and B Readers (“B” vs AI)

Finding	N	Threshold	Pairwise Agreement		Cohen’s Kappa	
			“B” vs “B”	“B” vs AI	“B” vs “B”	“B” vs AI
Abnormalities	1,569	0.50	0.9222	0.8995	0.8294	0.7767
Tuberculosis	1,264	0.50	0.9601	0.9318	0.9185	0.8627
Opacity[#]	1,443	0.50	0.9301	0.8950	0.8551	0.7759

[#]Opacity combines small and large opacities

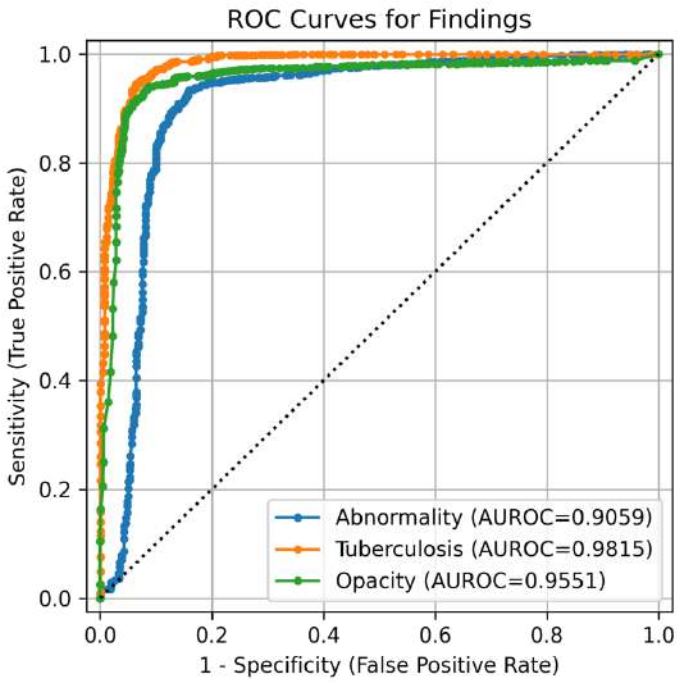
For measuring the diagnostic performance of each disease annotation, criteria such as Sensitivity, Specificity, Positive Prediction Rate (PPR), and Negative Prediction Rate (NPR) are utilized. These metrics are evaluated using the diagnostic threshold specified by the manufacturer, along with the area under the ROC curve.

Table 5 Diagnostic Performance of Each Finding by the System Compared to B Readers

Finding	N	Threshold	Sensitivity	Specificity	PPV	NPV	AUROC
Abnormalities	1,569	0.50	0.9369	0.8303	0.9107	0.8768	0.9059
Tuberculosis	1,264	0.50	0.9723	0.8873	0.9043	0.9660	0.9814
Opacity[#]	1,443	0.50	0.9653	0.7917	0.8728	0.9391	0.9551

[#]Opacity combines small and large opacities

Figure 1 ROC Curves Illustrating Diagnostic Performance for Each Finding



Analysis of Results

According to Table 6, when comparing Pairwise Agreement and Cohen's Kappa between B Readers and Qure.ai qXR ("B" vs AI) and among B Readers themselves ("B" vs "B"), Qure.ai qXR demonstrates performance close to that of B Readers (with a difference of less than 5%). For abnormalities, the agreement of among B readers scored higher than the agreement of each B reader and Qure.ai qXR by 2.27% (N=1,569). For tuberculosis, the agreement of among B readers scored higher than the agreement of each B reader and Qure.ai qXR by 2.83% (N=1,264). For opacity, the agreement of among B readers scored higher than the agreement of each B reader and Qure.ai qXR by 3.51% (N=1,443).

Table 6 Differences between Pairwise Agreement and Cohen's Kappa

Finding	Pairwise Agreement			Cohen's Kappa		
	B vs "B"	"B" vs AI	Diff	"B" vs "B"	"B" vs AI	Diff
Abnormalities	0.9222	0.8995	-2.27%	0.8294	0.7767	-5.27%
Tuberculosis	0.9601	0.9318	-2.83%	0.9185	0.8627	-5.58%
Opacity[#]	0.9301	0.8950	-3.51%	0.8551	0.7759	-7.92%

[#]Opacity combines small and large opacities

Regarding the lung tuberculosis screening, Qure.ai qXR, when analyzed on Dataset 1A2A–QureAI (300 images), showed diagnostic performance closely comparable to that of B Readers. It achieved an area under the receiver operating characteristic curve (AUROC) of 0.9841, sensitivity of 0.9723, and specificity of 0.8873 at a threshold of 0.50.

Referring to The Target Product Profiles (TPPs) for a rapid non-sputum-based biomarker test for tuberculosis detection by the World Health Organization (WHO), as shown in Table 7, it can be observed that each test scenario has different criterions for sensitivity and specificity.

Table 7 TPP for a rapid non-sputum-based biomarker test for tuberculosis detection

	Minimal Requirements		Optimal Requirements	
	Sensitivity	Specificity	Sensitivity	Specificity
Smear-replacement test	Overall >80% Positive >99% Negative >60%	98%	Overall >95% Positive >99% Negative >68%	98%
Non-sputum based biomarker test	Overall >65% Positive >98%	98%	Positive >98% Negative >68%	98%
Triage test	90%	70%	95%	80%

Reference: https://academic.oup.com/jid/article/211/suppl_2/S29/2490781

The Minimal Requirements and Optimal Requirements in the WHO TPPs (Target Product Profiles) outline the minimum and ideal thresholds for sensitivity and specificity that such a test should meet.

The Minimal Requirements indicate the minimum acceptable level of sensitivity and specificity that the test should achieve to be considered effective for tuberculosis detection. These criteria serve as a baseline standard for performance.

The Optimal Requirements represent the desired ideal performance levels for sensitivity and specificity. Meeting or exceeding these requirements would indicate a highly accurate and reliable test for tuberculosis detection.

The results of tuberculosis screening using Qure.ai qXR at different thresholds compared to the WHO TPP criteria, with the highest threshold that yields the closest specificity to the WHO TPP, are presented in Table 8.

Table 8 Sensitivity and Specificity Values at Different Thresholds according to WHO TPP Criteria

Threshold	Sensitivity	Specificity
0.9350	0.7437	0.9800
0.5080	0.9723	0.8873
0.1050	0.9953	0.8006
0.0540	0.9984	0.7054

Furthermore, when comparing the results obtained with the WHO TPP criteria, it was found that Qure.ai qXR met the requirements for the Triage test (for both the Minimal Requirements and Optimal Requirements) and the Non-sputum based biomarker test (for the Minimal Requirements criteria). The test outcomes are summarized in Table 9.

Table 9 Results of Tuberculosis Screening by Qure.ai qXR
according to WHO TPP Criteria.

	Minimal Requirements	Optimal Requirements
Smear-replacement test	Not Pass	Not pass
Non-sputum based biomarker test	Pass	Not pass
Triage test	Pass	Pass

Supplementary Table

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve which can be used to visualize the performance of a classifier at various thresholds. By adjusting the threshold, one change the trade-off between sensitivity and specificity. Table S1 and S2 details different sensitivity and specificity values across varying classification thresholds for abnormalities and tuberculosis, respectively.

Table S1 Sensitivity and Specificity Across Varying Classification Thresholds for Abnormalities. (Manufacturer's recommended threshold value is 0.50)

Threshold	Sensitivity	Specificity
0.9760	0.0217	0.9800
0.9500	0.1689	0.9517
0.9250	0.3684	0.9352
0.9000	0.631	0.9223
0.8750	0.7731	0.9093
0.8520	0.8362	0.8952
0.8250	0.8706	0.8810
0.8030	0.8993	0.8598
0.7750	0.9165	0.8492
0.7500	0.9235	0.8445
0.7450	0.9242	0.8422
0.7190	0.9299	0.8422
0.6720	0.9312	0.8410

Threshold	Sensitivity	Specificity
0.6350	0.9344	0.8398
0.5320	0.9344	0.8327
0.3120	0.9382	0.8292
0.2750	0.9382	0.8257
0.2520	0.9395	0.8210
0.2290	0.9433	0.8174
0.2000	0.9509	0.7750
0.1750	0.9554	0.7409
0.1520	0.9624	0.6584
0.1250	0.9777	0.5206
0.1000	0.986	0.3628
0.076	0.9962	0.2085
0.052	0.9987	0.0506

Table S2 Sensitivity and Specificity Across Varying Classification Thresholds for Tuberculosis. (Manufacturer's recommended threshold value is 0.50)

Threshold	Sensitivity	Specificity
0.9750	0.3528	0.9991
0.9500	0.6543	0.9887
0.9280	0.7848	0.9757
0.9010	0.841	0.9671
0.8750	0.8703	0.9575
0.8520	0.8924	0.9558
0.8280	0.9043	0.9480
0.8090	0.9146	0.9437
0.7900	0.928	0.9428

Threshold	Sensitivity	Specificity
0.5080	0.9723	0.8873
0.4830	0.9763	0.8839
0.4560	0.9778	0.8830
0.4490	0.9786	0.8813
0.4070	0.9786	0.8761
0.3880	0.981	0.8761
0.3660	0.981	0.8735
0.3310	0.9842	0.8718
0.3010	0.985	0.8674

Threshold	Sensitivity	Specificity
0.7510	0.9335	0.9411
0.7410	0.9383	0.9359
0.6770	0.9525	0.9255
0.6510	0.9557	0.9159
0.6330	0.9581	0.9133
0.6020	0.9644	0.9073
0.5950	0.9644	0.9047
0.5520	0.9715	0.8995
0.5300	0.9723	0.8951

Threshold	Sensitivity	Specificity
0.2900	0.9858	0.8631
0.1770	0.9873	0.8310
0.1540	0.9889	0.8250
0.1480	0.9905	0.8241
0.1000	0.9953	0.7981
0.0860	0.9976	0.7825
0.0510	0.9984	0.6898
0.0250	0.9992	0.4307

Typically, at a lower threshold, the model has high sensitivity but lower specificity. This means it correctly identifies most of the positives but also produces more false positives. At a medium threshold, there's a balance between sensitivity and specificity, which might be a good choice depending on the context. At a higher threshold, the model has high specificity but lower sensitivity. This is suitable when you want to be very certain about the positives but risk missing some.

Choosing the optimal threshold depends on the specific requirements of the task at hand. In some applications, high sensitivity might be more important, while in others, high specificity may be preferred.

Important Note

Qure.AI's abnormalities include the following: Atelectasis, Blunted Costophrenic Angle, Cardiomegaly, Cavity, Consolidation, Degenerative Spine Conditions, Elevated Hemidiaphragm, Fibrosis, Hyper Inflation, Lung nodule malignancy, Nodule, Non-Aortic Calcification, Opacity, Pleural Effusion, Pneumothorax, Prominence in Hilar Region, Reticulonodular Pattern, Scoliosis, Tracheal shift, and Tuberculosis. Some of the abnormalities involve other thoracic structures or nearby anatomical features which are beyond the lung fields. In the context of CXR screening, it is advised to be aware of these to ensure appropriate triage and supporting mechanisms.

3th November 2023