JAMIA Open, 2025, **8(1)**, ooae151
https://doi.org/10.1093/jamiaopen/ooae151
Research and Applications

Research and Applications

Towards human-AI collaboration in radiology: a multidimensional evaluation of the acceptability of AI for chest radiograph analysis in supporting pulmonary tuberculosis diagnosis

David Hua, BEng (Hons), BLaws[1,2], Neysa Petrina, PhD[1], Alan J. Sacks, MBBCh[3], Noel Young, MD[4,5], Jin-Gun Cho, MD, PhD[4,5,6], Ross Smith, MBBS, MPhil (CS)[1], Simon K. Poon, PhD, MPH[1,5,]*

[1]School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia, [2]Sydney Law School, The University of Sydney, Sydney, NSW 2050, Australia, [3]Our Medical Radiology, Sydney, NSW 2065, Australia, [4]Lumus Imaging, Sydney, NSW 2000, Australia, [5]Western Sydney Local Health District, Sydney, NSW 2145, Australia, [6]Sydney Medical School, The University of Sydney, Sydney, NSW 2050, Australia

*Corresponding author: Simon K. Poon, PhD, MPH, Building J12/1 Cleveland St, Camperdown, Sydney, NSW 2006, Australia (simon.poon@sydney.edu.au)

Abstract

Objective: Artificial intelligence (AI) technology promises to be a powerful tool in addressing the global health challenges posed by tuberculosis (TB). However, evidence for its real-world impact is lacking, which may hinder safe, responsible adoption. This case study addresses this gap by assessing the technical performance, usability and workflow aspects, and health impact of implementing a commercial AI system (qXR by Qure.ai) to support Australian radiologists in diagnosing pulmonary TB.

Materials and Methods: A retrospective diagnostic accuracy evaluation was conducted to establish the technical performance of qXR in detecting TB compared to a human radiologist and microbiological reference standard. A qualitative human factors assessment was performed to investigate the user experience and clinical decision-making process of radiologists using qXR. A task productivity analysis was completed to quantify how the radiological screening turnaround time is impacted.

Results: qXR displays near-human performance satisfying the World Health Organization's suggested accuracy profile. Radiologists reported high satisfaction with using qXR based on minimal workflow disruptions, respect for their professional autonomy, and limited increases in workload burden despite poor algorithm explainability. qXR delivers considerable productivity gains for normal cases and optimizes resource allocation through redistributing time from normal to abnormal cases.

Discussion and Conclusion: This study provides preliminary evidence of how an AI system with reasonable diagnostic accuracy and a human-centered user experience can meaningfully augment the TB diagnostic workflow. Future research needs to investigate the impact of AI on clinician accuracy, its relationship with efficiency, and best practices for optimizing the impact of clinician-AI collaboration.

Lay Summary

Artificial intelligence (AI) technology has the potential to transform radiological practice and increase clinician accuracy and efficiency through its ability to triage patient cases and generate diagnostic recommendations. However, evidence for the impact of AI in real-world settings is limited as past studies have focused on its technical performance in highly controlled laboratory settings. Limited attention has been given to how it affects clinical decision-making and healthcare delivery outcomes. This study builds the evidence base for the real-world impact of AI by evaluating the diagnostic accuracy, clinical workflow implications, and task productivity consequences of implementing a commercial AI system called qXR for supporting Australian radiologists in diagnosing tuberculosis (TB). The results offer promising preliminary evidence that qXR performs comparably to human radiologists, optimizes resource allocation through redistributing time spent from normal to abnormal cases, and is regarded favorably by clinicians because of its human-centered user experience and minimal workflow disruptions. This research provides medical institutions with a blueprint for assessing the suitability of AI products for use in their TB diagnostic workflows and specific clinical context. This framework can be continually used in clinical AI monitoring systems to enable issue detection, performance maintenance, and long-term safety and quality assurance.

Key words: pulmonary tuberculosis; artificial intelligence; evaluation; human factors; translational impact; technical performance.

Introduction

The progress of the international community in ending tuberculosis (TB) as a global public health challenge by 2030 has been significantly disrupted in recent years due to the increasing incidence of TB infections and mortality, which is largely attributable to the debilitating impacts of the Covid-19 pandemic.[1] The proliferation of mature commercial AI-based computer-aided detection (CAD) products for chest radiology presents a timely opportunity to address this problem as their deployment is expected to improve the efficiency and accuracy

Received: February 21, 2024; **Revised:** October 25, 2024; **Editorial Decision:** December 17, 2024; **Accepted:** January 22, 2025

of TB diagnostic services. Indeed, the World Health Organization (WHO) has recommended the use of CAD systems as a supplement or alternative to human interpretation of chest X-rays (CXRs) for the screening and triage of TB in populations aged 15 and above.[2] This is because chest radiographs are a sensitive, cost-efficient imaging modality for diagnosing TB and often only require one medical image for successful detection.

In Australia, the integration of AI technology into radiological systems is still at a nascent stage.[3] The evidence base for its risk and impact in real-world contexts is markedly lacking as most studies have been conducted in simulated settings, which cannot capture the complexities of daily clinical practice. This may hinder efforts to develop the best practices necessary for ensuring the safe, responsible, and effective deployment of AI applications in radiology. Conducting comprehensive health technology assessments that move beyond diagnostic accuracy to consider human factors and translational impact will be critical to addressing this gap.[4] Thus, this study aims to evaluate the technical performance, usability and workflow aspects, and health impact of the implementation of a commercial AI system called qXR, developed by the Indian-based AI vendor, Qure.ai, for assisting human radiologists with diagnosing TB in CXRs within an Australian context. This is modeled after the AI evaluation framework established by the American Medical Informatics Association and the multiphase research framework proposed by Park et al., which provide methodological guidance for evaluating different phases of the health AI lifecycle.[5],[6]

This study is an industry-academia research collaboration with an Australian medical imaging service providers who operate clinical sites that provide TB screening services. They have implemented qXR with the goal of improving the efficiency and accuracy of their radiologists who are responsible for performing over 250 000 chest examinations annually. qXR was chosen as a suitable tool to aid radiologists with conducting TB diagnosis because of its ability to interpret medical images and detect classic and atypical radiographic manifestations of pulmonary TB. Its primary features are generating diagnostic recommendations based on a prespecified criteria of clinical features and case triaging, which enables suspected abnormal cases to be prioritized for quicker review.[7] It is intended to function as a diagnostic assistive and productivity enhancing tool, and it does not seek to replace the role of radiologists who have ultimate responsibility in making diagnostic decisions. qXR has been trained on a large, diverse dataset of over 4.4 million samples and has been tested in various high TB-prevalence, low-resource countries.[7] It has not been evaluated extensively in low TB-prevalence, high-resource countries, and to date, no clinical studies of qXR have been conducted in an Australian setting.[8],[9] Therefore, the application of qXR for TB diagnosis in this context is novel given Australia is a well-resourced country with a low TB prevalence rate of 5.6 cases per 100 000 people and a diverse population representing over 250 different ethnicities.[10] This health technology assessment of qXR was conducted independently for the medical imaging service provider without the financial involvement of Qure.ai.

Methods

A 3-stage evaluation roadmap was developed to assess the acceptability of AI from a technical, end user, and organizational perspective. It leverages a mixed-methods approach that synthesizes the multi-faceted insights of diagnostic accuracy, behavioral, and workflow measurements. This aims to provide medical institutions with a blueprint for extracting evidence to determine the maturity and suitability of AI for adoption and to inform AI policy and governance decisions.

Stage I—technical evaluation study

Assessing the technical performance of qXR on CXRs representative of the target population demographics is critical to detect bias in its design and training data and in turn inform bias mitigation measures for improving safety. While qXR has been trained on a globally diverse dataset, it cannot be guaranteed that its performance observed in other settings will be replicable here. A retrospective diagnostic accuracy evaluation was conducted on anonymized patient cases extracted from clinical sites that involved comparing the accuracy of qXR against a human radiologist specializing in TB diagnosis with over 25 years of experience and a reference standard of mycobacterial culture. A sample of 126 CXRs (64 cases without TB, 62 cases with TB) was used. The distribution of positive and negative cases does not reflect the background prevalence of TB in the underlying population as the incidence rate of abnormal cases would be very low, which would make it difficult to properly assess the accuracy of qXR on a smaller sample size. Hence, a roughly even split of cases was selected in a randomized, deidentified manner to obtain a more accurate image of qXR's diagnostic accuracy. Summary statistics were generated to facilitate a pairwise comparison of qXR, the radiologist, and the reference standard. This will determine if qXR has an acceptable level of accuracy by satisfying the WHO's minimum threshold values of 0.9 for sensitivity and 0.7 for specificity for CAD systems used in TB detection.[2] This will indicate whether qXR is comparable to human performance and has sufficient technical merit to be used as a diagnostic assistant. This process was informed by the Food and Drug Administration's guidelines on clinical evaluation of software as a medical device.[11]

Stage II—usability and workflow study

Investigating the clinician-AI interaction process is critical to contextualize and understand how qXR impacts task effectiveness, efficiency, and satisfaction.[12] A qualitative human factors evaluation was conducted to understand the extent to which radiologists found qXR acceptable for use in daily clinical practice. This involved examining the nature of the AI-assisted workflow, system usability, algorithm understanding and explainability, and practical implementation lessons. Workflow process analysis was conducted on ethnographic observation data to determine how the implementation of qXR affects the decision-making process and workflow practices of radiologists. This involved radiologists providing a real-time work demonstration in completing a patient case, starting with opening a patient case and ending with submitting a TB diagnostic report, accompanied by explanations of their actions and reasoning for them particularly for interactions with qXR. Post-observation questions were asked to clarify inconsistencies between observed recurring behavior across radiologists and the idiosyncratic habits of individual clinicians. A scientifically descriptive account of the AI-assisted workflow was generated from this and visually summarized using a task hierarchy diagram and process flowchart. This was guided by the Workflow Elements model, which provides a conceptual structure of what to consider when examining workflow practices in healthcare.[13] Thematic analysis was performed on semi-structured interview data to capture the end user experience of qXR in daily clinical practice. Emergent themes and subthemes were derived from at least 3 radiologists raising similar thoughts on a particular issue to ensure they were views reasonably representative of the participant population. Interview questions were modeled after the contents of the Theoretical Framework of Acceptability (TFA), which is a health psychology framework that offers a systematic approach to investigating the acceptability of digital healthcare interventions in terms of the extent to which clinicians consider it to be appropriate based on their cognitive and emotional responses to it.[14] Questions were further inspired by prior empirical studies employing the TFA to study the acceptability of other digital healthcare interventions (eg, telehealth, chatbots) and relevant sociotechnical factors (eg, system trust, social influence) extracted from the human-machine interaction literature.

The study inclusion criteria required that participants be a radiologist employed by the industry partner and have used qXR for at least 1 month to ensure they have spent meaningful time using it in daily clinical practice. Volunteer sampling was used to avoid radiologists feeling obligated to partake in the study since it was a collaboration with their employer. Six out of 8 radiologists in the organization's TB diagnosis program elected to participate; 5 were male and aged 65 or above while 1 was female aged between 35 and 44. Ethical approval for this study was received from the University of Sydney Human Research Ethics Committee (Project No. 2022/141). Data collection and analysis were handled by 2 researchers experienced in qualitative methodologies.

Stage III—health impact study

Assessing the translational impact of qXR is necessary to generate evidence for its purported value proposition in improving healthcare delivery outcomes. This study focuses on measuring the task productivity impact of qXR because of its significance to the deployment context where there is a high daily patient volume and as there is limited empirical research on the productivity effects of radiological AI products. Turnaround time metrics, which are widely used indicators in health economics studies and radiological organizations as they provide insight into work, financial, and patient satisfaction performance, will capture whether qXR acceptably improves workflow efficiency.[15]

Two turnaround time metrics were used to quantify how qXR impacts the time taken to process a patient case for normal vs abnormal cases for radiologists individually and collectively: (1) case turnaround time covering the time from when a radiographer takes a CXR of the patient to when a radiologist finalizes their diagnostic report; and (2) reporting turnaround time covering the time from when a radiologist opens a patient case to when they submit their report. This will be based on analyzing timestamp and audit log data, which is automatically collected by the industry partner's radiology information systems, using Python and SQL. These data cover 2 distinctive time periods reflecting the clinical environment pre-AI implementation (June 2021 to August 2021) and post-AI implementation (September 2021 to August 2022). Only 3 months of data prior to AI integration were available because of recent changes to the organization's digital healthcare infrastructure. The number of cases completed on a monthly basis across this timeframe was also collected to contextualize whether any productivity impacts were influenced by workload changes.

Results

Stage I—technical performance study

The accuracy evaluation results are summarized in Table 1. Comparing the performance of the radiologist against qXR, sensitivity was 0.95 vs 0.90 ($P = .25$; McNemar's test) while specificity was 0.81 vs 0.70 ($P = .023$) with a positive predictive value of 0.83 vs 0.75 and a negative predictive value of 0.94 vs 0.88. qXR is reasonably aligned with the radiologist (correlation = 0.8391) but less so for the reference standard (correlation = 0.6176). This means qXR satisfies the minimum performance thresholds specified by the WHO and is approaching human performance, but it is still considerably less accurate than microbiological tests. These results should be treated as promising yet inconclusive evidence of qXR being reasonably safe for use as a decision-support tool for TB screening in an Australian context.

Stage II—usability and workflow study

Understanding the nature of the AI-assisted workflow

Figure 1 illustrates the AI-assisted workflow that involves 3 stages including image processing, system triaging, and image analysis and reporting. First, the image processing stage concerns how the patient chest radiograph and secondary capture, and a duplicate image of a CXR superimposed with annotations containing qXR's diagnostic recommendations is stored in the picture archiving and communication systems

Table 1. Pairwise comparison of qXR, the radiologist, and the reference standard in detecting TB.

Pairwise comparison	Correlation	Sensitivity (recall)	Specificity	Precision (PPV)	NPV	Accuracy	*P*-value (McNemar's)
qXR vs radiologist	0.8391	0.9577 (0.8638-0.9847)	0.8727 (0.7552-0.9473)	0.9067 (0.8291-0.9511)	0.9231 (0.8216-0.969)	0.9134 (0.8503-0.956)	0.3428
qXR vs reference standard	0.6176	0.9032 (0.8012-0.9637)	0.7031 (0.5758-0.8109)	0.7467 (0.667-0.8125)	0.8824 (0.7752-0.9422)	0.8016 (0.7212-0.8673)	0.0164
Radiologist vs reference standard	0.7703	0.9516 (0.865-0.9899)	0.8125 (0.6964-0.8992)	0.8310 (0.7464-0.8915)	0.9455 (0.851-0.9813)	0.8810 (0.8113-0.9318)	0.0389

The reported metrics include a 95% confidence interval where applicable. PPV: positive predictive value; NPV: negative predictive value.

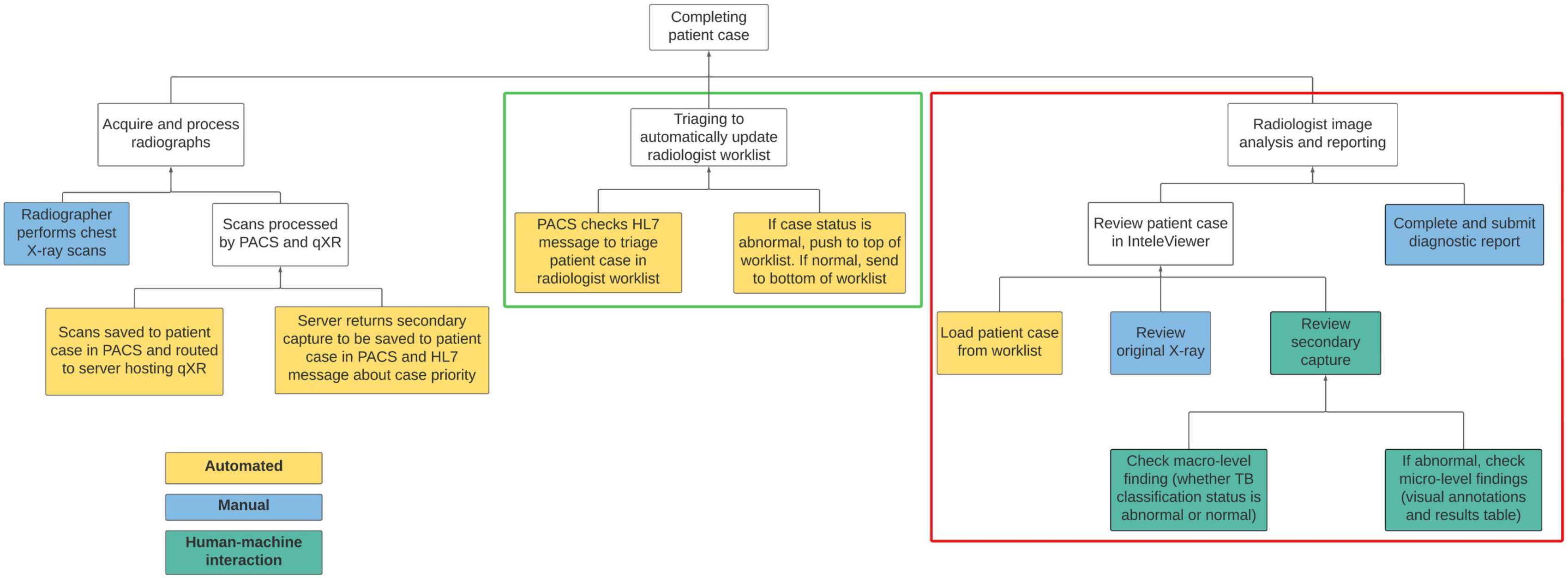
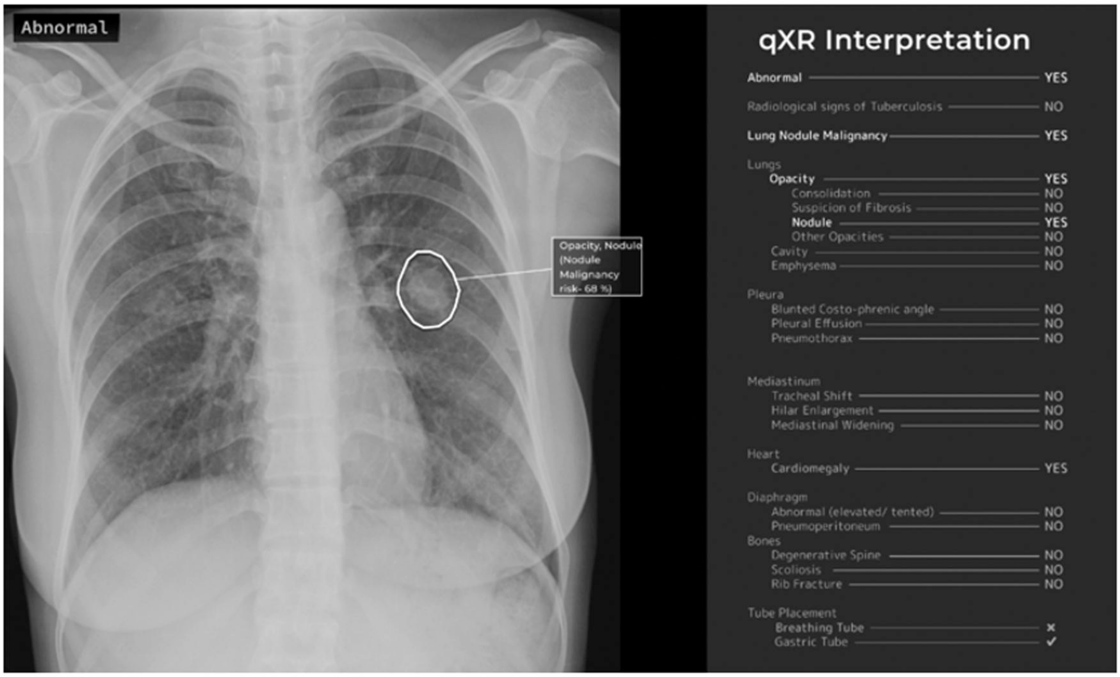

Completing patient case
Acquire and process radiographs
Radiographer performs chest X-ray scans
Scans processed by PACS and qXR
Scans saved to patient case in PACS and routed to server hosting qXR
Server returns secondary capture to be saved to patient case in PACS and HL7 message about case priority
Triaging to automatically update radiologist worklist
PACS checks HL7 message to triage patient case in radiologist worklist
If case status is abnormal, push to top of worklist. If normal, send to bottom of worklist
Radiologist image analysis and reporting
Review patient case in InteleViewer
Complete and submit diagnostic report
Load patient case from worklist
Review original X-ray
Review secondary capture
Check macro-level finding (whether TB classification status is abnormal or normal)
If abnormal, check micro-level findings (visual annotations and results table)
Automated
Manual
Human-machine interaction

Figure 1. Task hierarchy diagram outlining the steps of the AI-assisted workflow. The green rectangle is concerned with the triaging function of qXR while the red rectangle is associated with the diagnostic recommendation function.



qXR Interpretation

Abnormal	YES
Radiological signs of Tuberculosis	NO
Lung Nodule Malignancy	YES
Lungs	
Opacity	YES
Consolidation	NO
Suspicion of Fibrosis	NO
Nodule	YES
Other Opacities	NO
Cavity	NO
Emphysema	NO
Pleura	
Blunted Costo-phrenic angle	NO
Pleural Effusion	NO
Pneumothorax	NO
Mediastinum	
Tracheal Shift	NO
Hilar Enlargement	NO
Mediastinal Widening	NO
Heart	
Cardiomegaly	YES
Diaphragm	
Abnormal (elevated/ tented)	NO
Pneumoperitoneum	NO
Bones	
Degenerative Spine	NO
Scoliosis	NO
Rib Fracture	NO
Tube Placement	
Breathing Tube	✗
Gastric Tube	✓

Figure 2. Example of an annotated CXR produced by qXR containing diagnostic recommendations for TB. Adapted from Ref.[7]

(PACS) where clinicians can access them using diagnostic viewer software called InteleViewer. An example of a secondary capture generated by qXR is depicted in Figure 2. Second, the triaging stage captures how the patient worklist queue for radiologists is updated based on HL7 messages, which reflect the TB classification status issued by qXR. Lastly, the image analysis and reporting stage detail how radiologists interact with qXR to interpret medical images and complete their TB diagnostic reports. This workflow process is typically repeated several hundred times daily for each radiologist.

The impact of qXR on the clinical workflow arises during system triaging where it replaces the first-in-first-out approach to patient case management by prioritizing worklists based on suspected abnormalities, and during image analysis and reporting which involves human review of qXR's outputs to inform the diagnostic process. The decision-making process of the AI-assisted workflow is portrayed in Figure 3. When radiologists open patient cases from the top of their worklist queue, they are presented with the original CXR on the left and the secondary capture on the right. They first analyze the original CXR to form a preliminary TB diagnosis before verifying it against the binary classification given by qXR. If qXR has classified the case as abnormal radiologists will review its annotations to ensure they have not overlooked any flagged abnormalities and inform their diagnosis where relevant. Conversely, if qXR has classified, the case as normal radiologists will submit their diagnosis without adjustment.

Radiologist perspectives of using qXR

Table 2 presents the results of thematic analysis with representative quotations.

System usability

Radiologists expressed a high level of satisfaction with the usability of qXR. This was largely attributable to qXR causing minimal disruption to their well-established workflow habits and there being a justified yet limited increase in the workload burden. The adjustment process was considered simple and intuitive as it took very little time for radiologists to become accustomed to integrating the suggestions of qXR into their clinical judgement. There was no significant learning curve in the sense of having to undergo formal training to interact with new software, allowing them to immediately use it and quickly experience its benefits. Radiologists cite that qXR's outputs automatically loading and being embedded in InteleViewer as enabling this by eliminating unnecessary inconveniences associated with constantly repeat manual steps (eg, navigating between windows or applications, using a separate device) to access the secondary capture for each case, which would quickly become tiresome and frustrating given their high daily workloads. This meant the increased task burden of the AI-assisted workflow was limited to reviewing the suggestions of qXR and incorporating it into their TB diagnosis where relevant, rendering the time and effort required overall only slightly more demanding than the conventional workflow. Having a user-friendly, human-readable interface layout which contained a meaningful level of clinical detail yet could still be reviewed with seconds was important in achieving this. Radiologists stressed that avoiding an information overload which was time-consuming and cognitively taxing to navigate were critical to qXR having high usability.

Algorithm understanding and explainability

Radiologists reported having limited knowledge of how qXR operates yet did not believe this adversely affected their willingness to use qXR despite it being a black box system lacking explainability for how and why it reaches diagnostic conclusions. A common analogy raised was the fact they do not understand the technical mechanics behind how a motor vehicle functions but are still willing to use it provided it has been certified by authoritative bodies and they can see when

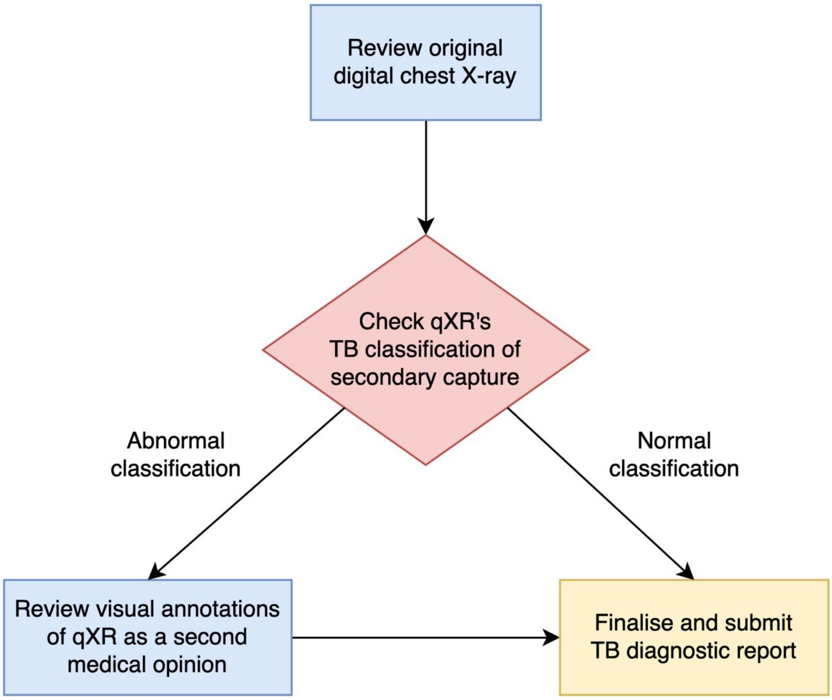

Review original digital chest X-ray
Check qXR's TB classification of secondary capture
Abnormal classification
Normal classification
Review visual annotations of qXR as a second medical opinion
Finalise and submit TB diagnostic report

Figure 3. Process map of the image interpretation and analysis stage of the AI-assisted workflow.

Table 2. Thematic analysis summary of the clinician end user experience.

Themes	Subthemes	Representative quotes
System usability	Learning process and transition from the conventional to the AI-assisted workflow was seamless	*"[Adjusting took] maybe 3-5 minutes . . . the training is close to nil because there is no training needed. It's just reading abnormal or not abnormal, seeing what [qXR] says and then you make your own assessment. I'm very happy with it [being a] digital helper."*
	Automatically loading qXR's outputs and embedding it in existing work software minimized workflow disruptions and inconveniences	*"If I had to scroll here, scroll there then that would be a pain. It's hands-off. It's good that the X-rays and secondary capture are presented side by side without me having to open and close another window."*
	qXR has a human-readable interface layout that is mindful of clinical informational needs and limited cognitive resource	*"[qXR] is clear and understandable to read . . . very self-explanatory. [It has] what you need, you don't need additional detailed information. [I get what I need] with a quick look."*
Implementation lessons	Lack of best practices to guide AI usage can create uncertainty and suboptimal usage	*"Nobody's told me what we are supposed to do with qXR. I'm only using it the way I think I should. If I get [guidelines] then I'll know if I'm using it correctly."*
	Restrictive policies around AI use which limit professional clinician autonomy may impair end user willingness to embrace it	*"Use it [and in this way] if you want, that's important to say. If [we are told] you must use [AI and in this way], everyone will say I'm never going to use it. Big cultural factor. The difference between must and can, the implementation is 0 or 100."*
	Time-efficient feedback processes are needed to sustain user contributions for long-term system improvement	*"A manual feedback form is so cumbersome. It has to be quick and seamless otherwise it won't be used . . . we are so busy, there's just so much to do."*
Algorithm explainability	Lack of explainability in qXR's outputs can create confusion for unexpected diagnoses	*"No explanation is given for how [qXR] interprets the cases and picks up findings . . . sometimes I struggle with understanding why it [gave a certain diagnosis]."*
	Radiologist willingness to engage with qXR is not hindered by its lack of explainability	*"[Using qXR] is like running a car really. The vast majority of times you don't need to know how the engine runs. Just drive it."*

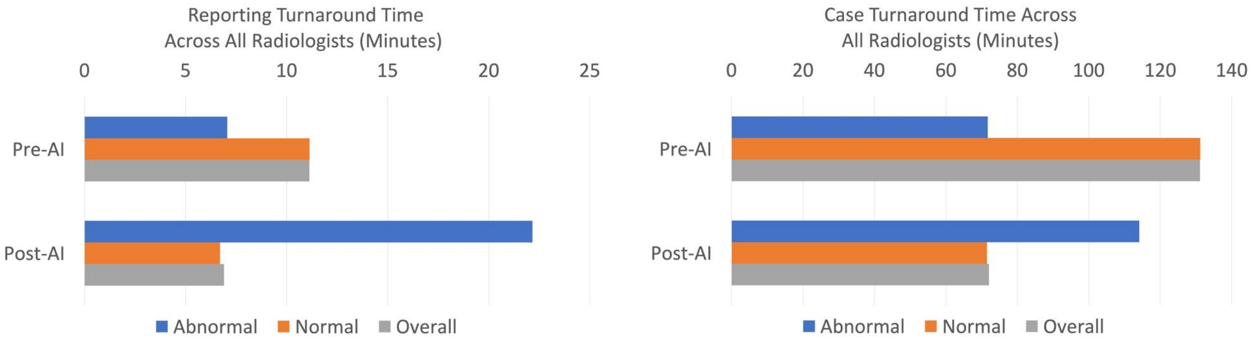


Figure 4. Mean case and report turnaround time before and after AI integration for all radiologists.

it works or fails by operating it. What mattered to radiologists was having the assurance that qXR has been tested as safe for clinical use (eg, by regulators, auditors, independent researchers) and it providing diagnostic recommendations which they can make sense of and are reasonably accurate. Hence, most radiologists believed that having algorithmic explainability (eg, heatmaps highlighting regions of severity, natural language explanations for how different factors contributed to a diagnosis) was not necessary for their workflow, which is relatively simplistic compared to other imaging modalities and diseases (eg, diagnosing cancer in MRIs). They however acknowledged the lack of explainability occasionally created confusion when it flagged an unexpected, non-obvious abnormality and commented that explanations could be useful to informing their clinical reasoning in challenging cases or where qXR conflicted with their medical judgement.

Implementation lessons

Radiologists were unclear as to the policies governing how qXR should be used (eg, handling suspected misdiagnoses for simple vs complex cases). Consequently, they commented on the utility of having a best practices guide for qXR based on the collective input and experiences of radiologists to promote greater consistency and certainty. They believed this would be particularly beneficial to radiologists who have not previously used qXR by providing a clear structure on optimally approaching the AI-assisted workflow. High-level guidelines endorsed by deploying organizations and professional bodies were also noted as important to reducing clinician anxieties around the risks of AI use.

Clinicians stressed the need to record cases of disagreement with qXR to secure high-quality training data for facilitating an ongoing feedback loop that maintains and improves system accuracy over time. However, they commented there was limited motivation to engage with an early version of the feedback process, which involved manually collating details of suspected misdiagnosis incidents in a word document. This was considered unreasonably onerous and time-consuming given their demanding workloads. They however were highly receptive to the idea and eventual implementation of a feedback mechanism embedded within InteleViewer that enables them to flag a case for later review with only one mouse click. This highlights the importance of having a convenient, time-efficient feedback loop to better ensure radiologists actively report disagreement incidents.

The institutional culture in terms of mandatory vs voluntary usage of qXR affected perceptions of system suitability and clinician willingness to adopt it. Radiologists found the encouraging attitude of the industry partner in framing qXR as a collaborative partner, which is respectful of their professional autonomy as the ultimate diagnostic decision-maker and that seeks to enhance their specialist capabilities, was pivotal in overcoming attitudes of AI hesitancy and skepticism. They commented compelled usage, combined with AI issuing prescriptive directions overriding their judgement, would have led to them rejecting or resisting the use of qXR.

Stage III—health impact study

Figure 4 depicts the mean case and reporting turnaround time aggregated across all radiologists before and after the implementation of qXR. Table 3 reports the average reporting turnaround time for individual radiologists with standard deviation values. These are based on 18 abnormal cases and 19 475 normal cases for the pre-AI stage, and 3764 abnormal cases and 296 471 normal cases for the post-AI stage.

For abnormal cases, the mean case and reporting turnaround time substantially increased by 42.42 and 15.1 minutes, respectively. This however was not uniformly experienced by all radiologists as 2 saw a time decrease for abnormal cases, 3 saw an increase, and the remaining did not have any relevant data available. For normal cases, the mean case and reporting turnaround time decreased considerably by 59.11 and 4.43 minutes, respectively. All clinicians experienced an appreciable reduction in the time spent on normal cases, ranging from a notable 25% to a significant 194% decrease (1.73-6.17 minutes). All radiologists saw an overall time decrease, which is expected given the overwhelming majority of cases were normal cases. This productivity impact arose in a context where aggregate caseloads steadily increased by a monthly average of 45 abnormal cases and 2219 normal cases as shown in Figure 5. This further highlights the benefits that can accrue from radiologists being able to resolve normal cases more quickly with the assistance of AI.

Discussion

Improving technical performance and mitigating against bias

The measured accuracy of qXR is comparable with the findings of recent evaluation studies despite most being conducted in high TB-prevalence, low-resource countries.[8,9]

Table 3. Mean reporting turnaround time with standard deviation values for individual radiologists before and after AI integration.

	Pre-AI reporting time (min)			Post-AI reporting time (min and % difference)		
Clinician	Abnormal	Normal	Overall	Abnormal	Normal	Overall
P1	7.22 ± 1.87	9.23 ± 8.51	9.23 ± 8.52	7.12 ± 4.6 (−1.4%)	6.15 ± 3.41 (−50.1%)	6.15 ± 3.40 (−33.37%)
P2	7.76 ± 1.04	13.16 ± 16.18	13.15 ± 16.17	8.91 ± 13.13 (+14.82%)	6.43 ± 2.95 (−104%)	6.47 ± 2.93 (−50.8%)
P3	N/A	8.43 ± 5.14	8.43 ± 5.14	7.03 ± 5.43	6.70 ± 1.87 (−25.82%)	6.70 ± 4.67 (−20.52%)
P4	N/A	12.71 ± 19.68	12.71 ± 19.68	8.55 ± 7.41	6.54 ± 7.18 (−194%)	6.56 ± 7.18 (−48.39%)
P5	5.27 ± 0.045	8.38 ± 13.5	8.36 ± 13.4	7.99 ± 15.28 (+51.61%)	6.16 ± 6.27 (−36.04%)	6.18 ± 6.43 (−26.08%)
P6	N/A	13.57 ± 14.43	13.57 ± 14.43	55.48 ± 359	8.29 ± 99.87 (−63.69%)	9.28 ± 111.9 (−31.6%)
P7	5.87	7.34 ± 5.14	7.34 ± 5.14	6.34 ± 2.01 (+8%)	5.58 ± 1.79 (−31.54%)	5.59 ± 1.79 (−23.84%)
P8	9.2	9.17 ± 2.87	9.17 ± 2.85	6.87 ± 2.83 (−25.32%)	6.24 ± 2.75 (−46.96%)	6.24 ± 2.75 (−31.85%)

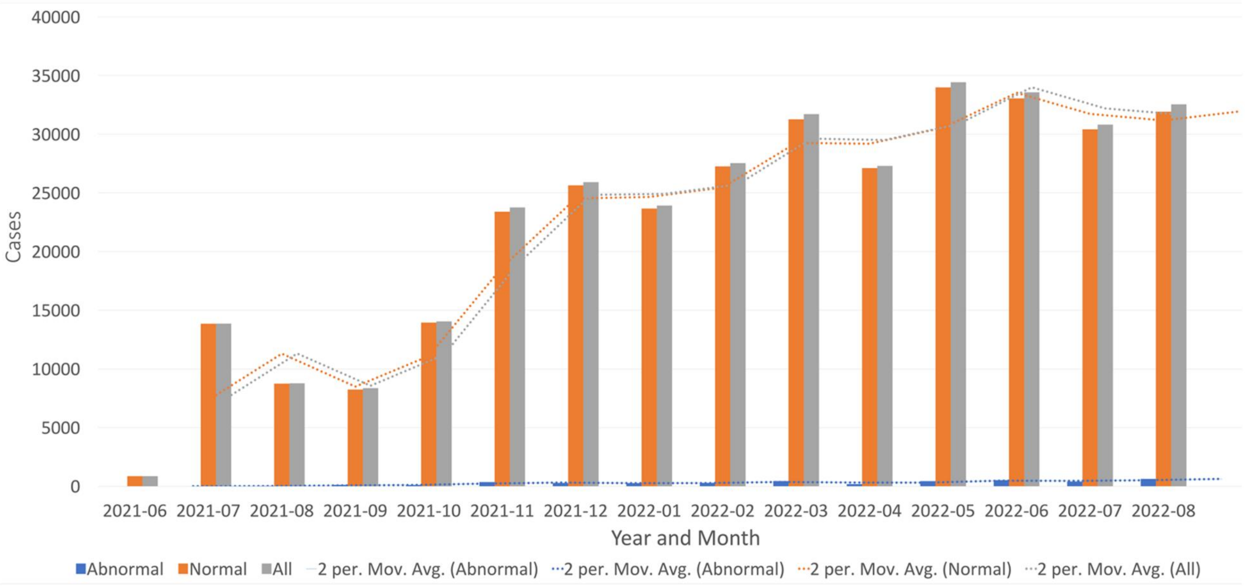


Figure 5. Case workload across time.

This is encouraging initial evidence that the observed performance of qXR in these settings generalizes reasonably well to a low TB-prevalence, high-resource environment with a multi-ethnic population. It however should be noted that since the diagnostic test accuracy evaluation was conducted on a CXR sample with an artificially higher prevalence of TB, the performance of qXR may have lower sensitivity and higher specificity when deployed in a population with a lower prevalence of TB. Further testing on a larger sample size of CXRs and with multiple radiologists is also needed to improve the statistical power and confidence in the results considering 126 CXRs is a modest figure.

Ongoing retraining of qXR with a focus on reducing false negatives is required given the considerable disparity between specificity and sensitivity. Further research is needed to understand the clinical characteristics and population-specific factors leading to undetected abnormalities. The fact that qXR correlates with the radiologist while the radiologist correlates with the reference standard has implications for how system retraining should be handled. Radiologists will be partly responsible for initiating the feedback loop process that enables continuous updating of qXR by reporting suspected misdiagnoses. If updated versions of qXR are trained based on feedback provided by a small number of radiologists, it may over time come to adopt their clinical approach and biases. This risks creating a self-reinforcing loop whereby the recommendations generated by qXR reflect the diagnostic patterns of certain clinicians which are reaffirmed in future feedback. This would arguably diminish its value in providing an alternative perspective that prompts clinicians to critically consider the clinical reasoning used by themselves and AI to reach an outcome. This is particularly problematic where qXR acquires unfavorable diagnostic habits, which cause it to consistently overlook certain clinical features and degrade its performance over time. Safeguarding against this demands verifying flagged misdiagnoses and conducting correlational analysis against microbiological tests or a diverse panel of radiologists with interrater reliability metrics for minimizing bias. Therefore, flagged cases of suspected misdiagnoses by qXR should be reviewed by at least 2 radiologists to reduce the risk of bias and error before it enters the feedback loop pipeline. Protocols should further be established to require additional human reviewers or a microbiological test for more complex cases where consensus cannot be reached. Statistical process control and quality management methods (eg, six sigma approach, power analysis) will also be needed to inform the minimum sample size used when continuously monitoring the accuracy of qXR.

The importance of a human-centered user experience to successful AI usage

The high usability of qXR was largely attributable to the limited and justified adjustment needed to transition from the conventional to AI-assisted workflow. This is consistent with status quo bias theory which posits that the introduction of new innovations that entail non-radical changes to working processes are more likely to be accepted by clinicians due to their preference to maintain well-established workflow habits.[16] This is supported by past research which has found that clinicians resist new health technologies that are needlessly disruptive to their workflow practices and do not conform to their mental models of clinical practice.[17] The compatibility of qXR with the existing workflow and the perception with which it has been designed to accommodate radiologist task requirements, rather than forcibly demanding their workflow to be structured around it, was a large incentive for user buy-in. Interestingly, the emphasis placed on qXR minimally increasing task burden is contrary to past usability research of healthcare IT, which suggests that clinicians are not overly concerned with ease of use as they pragmatically privilege factors that improve medical outcomes.[12,18] Clinicians with simpler workflows but greater caseloads, compared to those with more complex workflows but smaller caseloads, may value lower task burden more as any effort exerted on engaging with AI would be experienced more frequently. This was the case here as the sustained repetition of the workflow would amplify the increased effort and any negative experiences associated with qXR which could adversely impact occupational wellbeing and contribute to quicker fatigue and wariness. Yet, large proven improvements to clinical outcomes could be enough to outweigh user resistance associated with a more demanding and significantly altered workflow. This reflects the complex dynamic between affective attitude, burden, perceived effectiveness, and opportunity cost to the acceptability of AI technology as a digital healthcare intervention.[14]

The limited perceived value of algorithmic explainability diverges with past research which suggests it is an important factor affecting clinician trust, confidence, and satisfaction in AI.[19] This could be explained by the minimal reliance placed upon it by radiologists because of their decades of experience and the use case of qXR being a decision-support tool such that any harmful impacts are limited since the final diagnosis is handled by them. This may also be caused by lack of practical exposure to AI explainability techniques that would provide a point of comparison with the backbox nature of qXR's outputs. Radiologists concede the need for algorithm explainability would be more pronounced in settings with higher stakes, time criticality, or a more complex workflow (eg, operating semi-autonomously or autonomously, diagnosing a complex disease that has a high fatality rate). Future work should investigate the impact of different explainability techniques on system usability, workflow, and outcomes to determine which are most effective for chest radiology.

Healthcare organizations must be careful that their AI policies respect the professional autonomy of clinicians given its criticality to facilitating high satisfaction with AI.[12] Enactment of an institutionalized AI-assisted workflow which mandates or recommends certain conduct (eg, the order in which the original CXR and secondary capture should be viewed, how disagreement with AI should be handled) is crucial in enabling organizational maturity around AI usage by minimizing variability and suboptimal work habits. However, organizations must be mindful of ensuring the development of systematized workflows is clinician-driven, regularly updated as the understanding of AI improves, and does not unduly restrict freedom to maintain individualistic approaches to diagnosis. Medical organizations should ensure to frame the narrative of AI implementation as an effort to augment the diagnostic capabilities of clinicians where the role of AI is to be a support tool whose recommendations they may accept or reject as they see fit based on their clinical judgement.[12] This will help alleviate fears around it being a threat to their professional identities by being deployed to compete with and eventually replace them which can be detrimental to user acceptability.[12]

The human factors evaluation was cross-sectional rather than longitudinal in nature. Hence, it cannot capture how the clinician-AI dynamic might evolve over time as radiologists become more comfortable and proficient with using qXR and as system changes are made based on user feedback. Future work should be conducted across multiple time points to better capture different levels of system exposure and monitor for behavioral changes. The study demographics were largely uniform (male, aged 65 or above, over 20 years of experience) and therefore caution should be taken to generalizing the results to younger radiologists with limited experience. While this homogeneity will ensure less variability, future studies should ideally have diverse demographics.

The clinician-AI interaction insights presented here could be impacted by half the radiologists having no prior exposure to AI in clinical practice and the other half having previous experience with one AI system other than qXR. This could have affected their experiences of using qXR and in turn their perceptions of its utility and flaws. Those with prior experience of using AI might have a better practical understanding of qXR's capabilities and limitations. This could have enabled them to utilize qXR more effectively and confidently and may have made the learning and adjustment process for the AI-assisted workflow easier. Their favorable view of the usability and performance of qXR could also be magnified based on negative impressions of AI from past interactions with another AI tool. Meanwhile, those lacking past exposure to AI might have had more optimistic or skeptical expectations (eg, from independently reading academic literature or exposure to sensationalized headlines about medical AI technology) affecting their perceptions. However, the convergence of perspectives about the usability and impact of qXR suggests this was not a consequential factor affecting the results.

The implications of qXR facilitating a redistribution of time

The time improvement for normal cases is distributed across all radiologists rather than being concentrated among a subset. This demonstrates how an AI-assisted workflow can meaningfully improve productivity even for those with extensive diagnostic experience. The time savings accumulated across many normal cases in the medium to long term will be significant, enabling more cases to be processed in the same time and allowing more time to be allocated towards abnormal cases. This is important when considering that normal cases comprise an overwhelming 98.82% of the study sample of cases. Meanwhile, the increase in reporting time for abnormal cases is consistent with the anticipated benefits of using AI-driven triaging to ensure radiologists can review potentially abnormal cases more quickly and spend more time on them. In light of the growing workloads, this showcases how AI can optimize allocation of limited resources and improve the sustainability of healthcare delivery. However, this observed impact of AI-driven triaging emerges in a workflow context that exclusively concerns TB diagnosis such that qXR is only checking for TB even if other diseases may in fact be present. The effective priority of TB could be affected in clinical contexts where the medical imaging service provider has deployed the AI product(s) to triage for multiple diseases. Future studies should inquire into how the dynamics of AI-driven triaging in these scenarios could impact the turnaround time for the TB diagnostic workflow.

The redistribution of time from normal to abnormal cases has occurred despite a consistently growing workload. As most cases are normal, reflecting an imbalanced dataset and low incidence rate of TB, the productivity impact of qXR will necessarily be dominated by how normal cases are affected. If diagnostic accuracy is at least maintained, any time increase for abnormal cases is mitigated by them being a small proportion of all cases and the deployment setting being a non-critical, outpatient context. Nevertheless, it remains unclear whether radiologists are spending more time than justified on potentially abnormal cases such that significantly less time is allocated than warranted on normal cases which would be an excessive redistribution of time. This could manifest in the context of false negatives where clinicians overlook or are less thorough towards subtle abnormalities undetected by qXR, and in the context of false positives where clinicians search for non-existent or irrelevant abnormalities.

A historical analysis approach was taken to quantify the productivity impact of qXR by comparing turnaround time metrics before and after system deployment. The values reported in Table 3 reflect considerable variability in task efficiency across radiologists and between cases for each radiologist in terms of the pre-AI and post-AI reporting times. The impact on abnormal cases will be heavily skewed by the small sample size of 18 for the pre-AI stage which makes it unclear as to whether the time increase for abnormal cases reflects a meaningful effect or unrepresentative sample. This small sample size could explain the poor statistical precision in some of the results associated with unusually large standard deviation values. Another plausible explanation for the standard deviation values is that some participants work in a clinic environment where they are often interrupted when working on a case, prompting them to leave their workstation for an extended timeframe before they can return to complete it. For example, they might be interrupted to conduct interventional work or to attend a meeting. In the case of P6 who is a notable outlier based on significantly high post-AI reporting time values for abnormal cases, it is possible that they tend to leave InteleViewer with a patient case open while having a lunch break or after working hours which could artificially inflate their reporting time substantially. Because of resource limitations this study did not account for diagnostic correctness, including the level of agreement and disagreement between radiologists and qXR, which prevents an assessment being made as to whether the efficiency gains entail some trade-off with clinician accuracy. Understanding the extent to which radiologist accuracy is preserved, improved, or degraded with these efficiency gains and in what contexts is important to informing public health policy around when and how AI should be used. Future research should inquire into the impacts of AI on accuracy and its relationship with productivity by conducting randomized control trials where radiologists are randomly chosen to report with and without AI based on the time-to-correct diagnosis. It would also be worthwhile to granularly examine the impact of AI across varying categories of severity and clinical features associated with abnormalities to better understand its utility in different clinical scenarios.

Insights for medical institutions on selecting a suitable commercial AI system for integration into their TB diagnosis workflow

This evaluation study provides medical institutions with methodological guidance and structure around conducting assessments on the diagnostic accuracy, usability and workflow aspects, and healthcare delivery impact of commercial AI products for implementation in TB diagnostic workflows. By outlining a multi-phase evaluation roadmap for qXR, it highlights the high standard of evidence that should be gathered across technical, end user, and organizational dimensions during AI pilot trials to help determine if an AI product is suitable for a given clinical context. This is also useful in informing how clinical institutions view the evidence provided by other evaluation studies about AI products. For example, studies exclusively assessing the technical performance of an AI product might be useful to understand its diagnostic accuracy level but are insufficient to provide a complete picture about its acceptability for deployment due to the absence of insights on the clinician-AI interaction process and how patient outcomes are impacted.

This research further presented practical implementation and organizational governance lessons on integrating AI into the TB diagnosis workflow in a way that promotes safe and responsible usage, supports user uptake, and facilitates long-term system maturity and benefit realization. These lessons can assist clinicians with their transition to the AI-assisted workflow and better ensure organizations setup safeguards around AI usage including the establishment of feedback loop mechanisms to promptly manage adverse incidents and maintain AI performance across time. This is particularly significant in light of the WHO's warning that rapid AI adoption without rigorous ongoing testing and robust governance could cause medical errors that disrupt workflows and inflict patient harm at scale, thereby undermining trust in AI and delaying the long-term benefits of its use.[20]

This study however did not inquire into the diagnostic accuracy of clinician-AI collaboration and how clinical decision-making in the TB diagnostic workflow is affected by variations in key input factors. Consequently, there is limited understanding around how diagnostic decisions and outcomes are affected by: the influence of second medical opinions by human radiologists to resolve conflicts with AI; the weighting given to false negatives by AI and the risk of it contributing to human error; the patterns of TB presentation detected by AI; and the role of cut-off thresholds when set to default baseline levels by AI vendors or calibrated in an attempt to better reflect patient demographics. These are critical insights which would be valuable to assist medical organizations in better understanding the suitability of an AI product and how to best integrate it into their workflow, and should therefore be investigated in future research.

Conclusion

This research presents a multidimensional evaluative approach that assesses technical, end user, and organizational concerns to determine the acceptability of AI for adoption and inform AI policy and governance decisions in healthcare. When applied to a case study concerning an AI decision-support tool for assisting with TB diagnosis, it successfully generated promising preliminary evidence of how the AI system had reasonable accuracy, a human-centered user experience, and materially improved workflow efficiency. This framework can be repeatedly used as part of a clinical AI monitoring system to extract the data needed to facilitate issue detection (eg, sudden significant variability in outcomes) and system maintenance for long-term safety and quality assurance. The measurements used can easily be adjusted to accommodate other AI solutions and clinical tasks. Further research applying this methodology in varying clinical contexts is required to expand the evidence base around the impact of AI on health outcomes and to develop best practices for addressing unintended consequences and long-term sustainable process improvement to maximize the value of clinician-AI collaboration.

Author contributions

David Hua (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Neysa Petrina (Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing—review & editing), Alan J. Sacks (Formal analysis, Investigation, Resources, Writing—review & editing), Noel Young (Resources, Writing—review & editing), Jin-Gun Cho (Resources, Writing—review & editing), Ross Smith (Writing—review & editing), and Simon K. Poon (Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing—review & editing)

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data collected and analyzed in this study are not available because of ethical and privacy reasons.

References

1. World Health Organization. Global tuberculosis report; 2022. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022
2. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 2: screening—systematic screening for tuberculosis disease; 2021. https://apps.who.int/iris/bitstream/handle/10665/340255/9789240022676-eng.pdf
3. Royal Australian and New Zealand College of Radiologists. Artificial intelligence: the state of play 2019; 2019. https://www.ranzcr.com/college/document-library/artificial-intelligence-the-state-of-play-2019
4. Hua D, Petrina N, Young N, Cho JG, Poon SK. Implementing AI-based computer-aided diagnosis for radiological detection of tuberculosis: a multi-stage health technology assessment. In: *2023 IEEE International Conference on Digital Health (ICDH)*, 2-8 July 2023. IEEE; 2023: 353-355. https://doi.org/10.1109/ICDH60066.2023.00059
5. American Medical Informatics Association. AMIA 2024 artificial intelligence evaluation showcase. Accessed 19 February, 2024. https://amia.org/education-events/amia-2024-artificial-intelligence-evaluation-showcase
6. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open*. 2020;3:326-331. https://doi.org/10.1093/jamiaopen/ooaa033
7. Qure.ai. AI for blazing fast reporting on chest X-rays. Accessed 7 August, 2024. https://qure.ai/product/qxr/
8. Harris M, Qi A, Jeagal L, 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: e0221339. https://doi.org/10.1371/journal.pone.0221339
9. Hua D, Nguyen K, Petrina N, et al. Benchmarking the diagnostic test accuracy of certified AI products for screening pulmonary tuberculosis in digital chest radiographs: preliminary evidence from a rapid review and meta-analysis. *Int J Med Inform*. 2023;177:105159. https://doi.org/10.1016/j.ijmedinf.2023.105159
10. New South Wales Ministry of Health. Communicable diseases weekly report, 2022. https://www.health.nsw.gov.au/Infectious/Reports/Publications/cdwr/2022/cdwr-week-11-2022.pdf
11. United States Food and Drug Administration. Software as a Medical Device (SAMD): clinical evaluation—guidance for industry and Food and Drug Administration staff; 2017. https://www.fda.gov/media/100714/download
12. Hua D, Petrina N, Young N, Cho J-G, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. *Artif Intell Med*. 2024;147:102698. https://doi.org/10.1016/j.artmed.2023.102698.
13. Unertl K, Novak L, Johnson K, Lorenzi N. Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review. *J Am Med Inform Assoc*. 2010;17:265-273. https://doi.org/10.1136/jamia.2010.004333
14. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. *BMC Health Serv Res*. 2017;17:88. https://doi.org/10.1186/s12913-017-2031-8

15. Ozkaynak M, Unertl K, Johnson S, Brixey J, Haque SN. Clinical workflow analysis, process redesign, and quality improvement. In: *Clinical Informatics Study Guide: Text and Review*. Springer; 2022: 103-118.
16. Kim H-W, Kankanhalli A. Investigating user resistance to information systems implementation: a status quo bias perspective. *MIS Quarterly*. 2009;33:567-582.
17. Prakash AV, Das S. Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: a mixed-methods study. *Inform Manag*. 2021;58:103524. https://doi.org/10.1016/j.im.2021.103524.
18. Holden RJ, Karsh B-T. The technology acceptance model: its past and its future in health care. *J Biomed Inform*. 2010;43:159-172. https://doi.org/10.1016/j.jbi.2009.07.002.
19. Calisto FM, Santiago C, Nunes N, Nascimento JC. BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions. *Artif Intell Med*. 2022;127:102285. https://doi.org/10.1016/j.artmed.2022.102285.
20. World Health Organization. WHO calls for safe and ethical AI for health. Accessed 16 August, 2024. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health

JAMIA Open, 2025, 8, 1–12
https://doi.org/10.1093/jamiaopen/ooae151
Research and Applications