JURNAL LOCUS: Penelitian & Pengabdian
Volume 5 Number 01, Januari 2026
E-ISSN 2829-7334| P-ISSN 2829-5439
Hompage: https://locus.rivierapublishing.id/index.php/jl

Accuracy of Artificial Intelligence in Detecting Tuberculosis from Chest X Ray: A Systematic Review and Meta-Analysis of Diagnostic Performance

Singgih Priyambodo*, Iwan Danardono
Prembun Regional General Hospital, Indonesia
Email: singgihsinpe@gmail.com*

Abstract:
Tuberculosis (TB) remains a major global public health challenge, particularly in low-resource countries where access to trained radiologists is limited, making chest X-ray (CXR) screening difficult to scale. The advancement of artificial intelligence (AI) and computer-aided detection (CAD) technology offers a potential solution by providing automated TB detection and supporting diagnostic workflows. To assess their clinical readiness, this systematic review and meta-analysis was conducted using the PRISMA 2020 protocol and included studies from PubMed, Scopus, and Semantic Scholar that evaluated AI-CAD systems (index test) against microbiological or extended reference standards (reference standard). The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2 and QUADAS-C) tools were applied to measure risk of bias, and a random-effects model was used to estimate pooled diagnostic odds ratio (DOR). Six studies with approximately 38,940 participants were eligible for analysis. Results showed a pooled DOR of 0.133 (95% CI: 0.047–0.377), indicating a significantly lower diagnostic error rate (P=0.000). Although sensitivity was consistently high (83.3%–100%), specificity varied widely (26.8%–98.9%), resulting in notable heterogeneity and a wide prediction interval (0.003–6.411). These findings conclude that AI-CAD tools demonstrate strong potential for TB screening but should undergo local validation, threshold calibration, and operational evaluation before broad clinical implementation, especially where specificity remains below the WHO Target Product Profile.

Keywords: Tuberculosis, Artificial Intelligence, Computer-Aided Detection

Corresponding: Singgih Priyambodo
E-mail: singgihsinpe@gmail.com

INTRODUCTION

The persistent global public health threat posed by Pulmonary Tuberculosis (PTB) necessitates the continuous development of efficient and scalable diagnostic strategies (Marais et al., 2005). While gold-standard microbiological tests offer high accuracy, their utility in resource-limited, high-prevalence settings is severely restricted by infrastructure requirements and lengthy turnaround times (World Health Organization, 2023). Chest X-ray (CXR) screening is a critical, low-cost component of many diagnostic pathways; however, the shortage of trained human readers often creates bottlenecks and introduces subjectivity in interpretation (World Health Organization, 2023).

In response to these challenges, deep learning-based Artificial Intelligence (AI) algorithms have been developed to automate the detection of TB-consistent lesions on CXR images (Qin et al., 2021). These Computer-Aided Detection (CAD) systems, such as qXR and CAD4TB, are designed to provide rapid, objective scoring, thereby enhancing screening throughput and potentially decreasing reliance on specialized personnel (Marais et al., 2005).

Despite initial demonstrations of impressive performance metrics, concerns regarding the external validity, generalizability, and consistency of these AI algorithms persist (Dujardin et

al., 2023). Diagnostic accuracy studies (DTA) often show marked variation in reported sensitivity and specificity, reflecting differences in patient populations, imaging hardware, and the specific version of the AI software employed (Monu et al., 2020). A synthesized quantitative review is crucial not just to calculate the average effectiveness but, more importantly, to dissect and explain the profound variability observed in real-world deployment.

Understanding the determinants of heterogeneity is necessary to guide policy decisions, especially concerning the potential failure of these tools in new, unstudied environments, as suggested by the widely varying results (Page et al., 2021). This review therefore aims to provide a robust summary of diagnostic performance while critically investigating the factors that influence the reproducibility of AI-CAD systems.

The objectives of this systematic review and meta-analysis, formulated using the PICO framework (Population, Index Test, Comparator, Outcome), are: To determine the overall diagnostic accuracy of AI-based CAD tools for detecting Pulmonary TB from CXR images, utilizing the Diagnostic Odds Ratio (DOR) as the primary measure of effect. To quantify the extent of statistical heterogeneity and assess its clinical relevance by reporting the 95% Prediction Interval (PI) for the pooled effect. To explore potential sources of observed heterogeneity using subgroup analyses, specifically focusing on the impact of the type of reference standard (microbiological vs. non-microbiological) and the specific AI tool version used.

METHOD

The methodology employed in this systematic review and meta-analysis was defined a priori in a formal protocol, ensuring methodological transparency (Page et al., 2021). No substantive deviations or amendments were made to the pre-specified search strategy, eligibility criteria, or outcome measures following registration (Page et al., 2021).

Studies were selected based on the following pre-specified PICO-DTA criteria (Page et al., 2021): 1) Population: Eligible studies included human participants, encompassing adolescents and adults, undergoing evaluation or screening for PTB using a Chest X-ray. The population could represent either mass screening cohorts or symptomatic individuals presenting for diagnosis. 2) Index Test: The index test was defined as any AI-CAD system, employing machine or deep learning, designed for the automated detection or scoring of PTB findings on standard CXR images (e.g., qXR, CAD4TB). 3) Reference Standard: Inclusion required studies to use a robust reference standard. The primary standard accepted was microbiological confirmation, including culture, GeneXpert MTB/RIF, PCR, or TrueNat (Page et al., 2021). Studies using a non-microbiological reference standard (Non-MRS), such as radiological consensus or clinical proxy, were included for qualitative review and heterogeneity analysis but were specifically flagged as a potential source of verification bias. 4) Study Design: Only diagnostic accuracy studies (DTA) reporting sufficient data to construct 2X2 contingency tables (or equivalent metrics like sensitivity and specificity) were eligible for the meta-analysis. Reports had to be available as full texts in the English language and published in peer-reviewed journals or academic pre-print servers (Page et al., 2021).

Information Sources and Search Strategy: 1) Information Sources: A systematic search was executed across three major academic databases to identify relevant literature: PubMed (MEDLINE), Scopus, and Semantic Scholar (Page et al., 2021). The search spanned from the inception of each database up to. The date of the last search was. 2) Search Strategy: The full search strategy, including line-by-line commands and applied filters, is detailed in Supplementary Appendix A. The strategy systematically combined terms representing the Population ("tuberculosis," "TB"), the Index Test ("artificial intelligence," "machine learning," "computer-aided detection"), and the Diagnostic Method ("chest X-ray," "CXR") using validated Boolean logic (World Health Organization, 2023). Key limits applied included restricting results to human studies (Page et al., 2021). The exhaustive reporting of the search strategy is intended to ensure reproducibility, aligning with quality standards for systematic review methodology (Rethlefsen et al., 2021).

Study Selection and Data Collection Process: Selection Process: The selection process was conducted independently by two reviewers (R1 and R2) to minimize selection bias (Page et al., 2021). Initial screening involved title and abstract review to identify potentially relevant records. Subsequently, full-text reports were retrieved and assessed against the eligibility criteria. Disagreements between reviewers at any stage were resolved through discussion to reach consensus, or by consultation with a senior reviewer. 2) Data Collection: Data extraction was performed in duplicate by the two independent reviewers using a pre-defined and piloted extraction form (Page et al., 2021). Attempts were made to collect cell counts necessary for 2X2 contingency tables (TP, FP, TN, FN) and crucial descriptive data. Where necessary, study investigators were contacted to clarify reported results or to provide missing summary statistics (Page et al., 2021).

Data Items; 1) Outcome Measures: The primary measure of interest was the study-specific and pooled Diagnostic Odds Ratio (DOR), calculated from the 2x2 data. Secondary measures included sensitivity, specificity, and the Area Under the Curve (AUC). 2) Covariates: Detailed variables were sought to characterize the studies and explore heterogeneity, including 3: 3) Technical Data: Name and specific version of the AI-CAD tool (e.g., qXR v3.0, CAD4TB v7). 4) Methodological Data: Specific microbiological test used as the reference standard (e.g., Culture, GeneXpert). 4) Contextual Data: Geographical location, type of population screened (general population vs. symptomatic cases), and prevalence of TB. **5)** Administrative Data: Sources of funding or support for the primary study.

The Quality Assessment of Diagnostic Accuracy Studies Version 2 (QUADAS-2) tool was used to assess the methodological quality and risk of bias (RoB) in all included studies (Whiting et al., 2011). Given the comparative nature of AI validation studies (often comparing AI to a human reader or standard test), the QUADAS-C extension was applied to evaluate potential comparative biases (Ferreira-González et al., 2022).

RoB Domains and AI Context: Judgments of RoB were made across four domains: Patient Selection, Index Test, Reference Standard, and Flow and Timing (Whiting et al., 2011). Particular attention was paid to the Index Test domain, where a high risk was assigned if the selection of the diagnostic threshold was optimized solely to maximize a single metric (e.g., sensitivity) without considering its downstream consequence (low specificity) (World Health

Organization, 2021). For instance, studies reporting specificities below 35% (Page et al., 2021) suggest a strong threshold bias intended for rule-out scenarios, which compromises the test's utility for generalized application. Similarly, the Reference Standard domain was rated high risk if non-microbiological methods, such as consensus of unconfirmed readers, were used as the gold standard, raising concerns about verification bias (Page et al., 2021).

Applicability Assessment: Applicability concerns were assessed based on whether the population and the execution of the index test aligned with the review question. Concerns were raised if the study population was highly specific (e.g., only confirmed TB cases) or if the AI test was not deployed under real-world operational conditions, potentially limiting the generalizability of the findings to routine screening practice (Dujardin et al., 2023).

Data Synthesis and Statistical Analysis 1) Effect Measures: The Diagnostic Odds Ratio (OR) was used for pooling, calculated as .This metric provides a succinct summary of the test's overall ability to discriminate between true positives and negatives. 2) Statistical Model: Due to the expected and confirmed clinical and methodological variability (heterogeneity) among the included studies (Monu et al., 2020), a random-effects meta-analysis model was chosen to pool the log ORs (Glas et al., 2003). This model assumes that the true effect size varies across different study contexts. The analysis employed the DerSimonian and Laird method for estimating the between-study variance. The analysis was performed using (DerSimonian & Laird, 1986). Although the gold standard for DTA synthesis involves joint modeling of sensitivity and specificity (e.g., bivariate or HSROC models) to account for threshold effects (Reitsma et al., 2007), the pooling of the OR was justified as a necessary, robust simplification given the data structure reported in the primary studies, which often focused on a single summary estimate relative to a comparator. 3) Heterogeneity Assessment: Heterogeneity was assessed statistically using the I2 statistic and Cochran's Q test (Page et al., 2021). Critically, the 95% Prediction Interval (PI) was calculated. The PI provides a prospective range, indicating where the true effect of AI performance would likely lie if the tool were deployed in a new, comparable setting (Page et al., 2021).

Exploration of Heterogeneity: Two primary sources of heterogeneity were explored through subgroup analysis: Reference Standard Type: Comparing the pooled OR derived from studies using the rigorous Microbiological Reference Standard (MRS) versus those using the Non-Microbiological Reference Standard Non-MRS) (Page et al., 2021). AI Tool Platform: Comparing the performance of different AI tools and specific software versions (e.g., qXR versions vs. CAD4TB versions). Sensitivity Analyses: Sensitivity analyses were performed to confirm the robustness of the primary meta-analytic result (Page et al., 2021). These included removing studies with a high overall risk of bias (QUADAS-2/C assessment) and excluding the study that reported an effect estimate (OR = 1.309) suggesting lower accuracy than the comparator (Page et al., 2021).

RESULTS AND DISCUSSION

Study Selection

The systematic search yielded [Placeholder: X] unique records. Following screening of titles and abstracts, reports were excluded, and full texts for [Placeholder: Z] reports were

retrieved for detailed assessment. Ultimately, six studies met all pre-specified eligibility criteria and were included in the quantitative meta- analysis (Page et al., 2021). The selection process is detailed in the PRISMA 2020 Flow Diagram (Figure 1, Supplementary Material).

a) Excluded Studies

Reports excluded at the full-text stage primarily lacked appropriate microbiological reference standards or presented internal validation results unsuitable for calculating an external diagnostic accuracy measure (Page et al., 2021).

Characteristics of Included Studies

The included six studies collectively contributed data from a large population, involving an estimated 38,940 individuals undergoing TB evaluation (Page et al., 2021). The studies were geographically diverse, focusing on settings such as Lima, Peru, Ethiopia, and large screening cohorts in Bangladesh and India (Page et al., 2021).

The table summarizes the accuracy of various Computer-Aided Detection (CAD) tools in diagnosing tuberculosis (TB) from chest X-ray (CXR) images, comparing their performance with reference standards and radiologist evaluations. Here's the breakdown of the data:

1. Biewer et al., 2024: The study involved 1006 patients at a hospital in Lima, Peru, comparing two versions of qXR (v3.0 & v4.0) against culture and Xpert MTB/RIF tests. The sensitivity of the CAD tool was 0.91, and specificity was 0.32 for v4.0 with culture as the reference standard. However, the accuracy of the radiologists was not reported.
2. Binegdie et al., 2025: This research analyzed 3351 chest X-rays from adults in Ethiopia, with 552 images included in the study. It used qXR and GeneXpert MTB/RIF as the reference. The CAD tool achieved a sensitivity of 100% and specificity of 98.9%. The radiologists agreed with 94.4% of the AI results, confirming the AI's high performance.
3. Byrne et al., 2024: Conducted with 90 participants from a Primary Health Center in Karnataka, India, the study tested the CAD4TB tool against culture or PCR reference standards. The CAD tool had a sensitivity of 83.3% and specificity of 69%. Radiologists flagged 26 cases (29%) as probable or possible, with 60% of those confirmed by further testing.
4. Qin et al., 2021: This large study in Bangladesh included 23,954 individuals aged 15 years and above, using qXR v3, CAD4TB v7, and other tools for comparison against Xpert MTB/RIF. The qXR had an area under the curve (AUC) of 90.8% and specificity of 74.3% at 90% sensitivity, meeting WHO Target Product Profile (TPP) standards. The radiologists' performance was slightly lower, with specificity at 88.9% and lower predictive values.
5. Smriti et al., 2024: This study involved 1278 CXR pairs (digital and photo format) from Bihar, India, and evaluated the performance of qXR v3.2 against a consensus of 3 radiologists. The positive predictive agreement (PPA) was 92.22%, and the negative predictive agreement (NPA) was 82.08%. However, the radiologists were used for comparison, not as an independent reference.
6. Vijayan et al., 2023: Focused on 10,481 presumptive TB cases from informal providers in Nagpur, India, this study used qXR and microbiological (TrueNat) and clinical

evaluations as the reference. The CAD tool showed a sensitivity of 99.1% and specificity of 26.8%. The performance of the radiologists was not explicitly reported.

This table highlights the varying levels of sensitivity, specificity, and agreement between AI tools and radiologists, emphasizing the potential and limitations of CAD tools in TB detection based on CXR images across different regions.

The dominant AI platforms evaluated were qXR and CAD4TB, with several studies (e.g., Biewer et al., 2024; Qin et al., 2021) assessing different versions of these tools, illustrating the rapid technological advancement in this domain (Page et al., 2021). Five studies utilized a rigorous Microbiological Reference Standard (MRS), including GeneXpert MTB/RIF, culture, or TrueNat. One study employed a Non-Microbiological Reference Standard (consensus of three radiologists) (Page et al., 2021). The specific characteristics are summarized in Table 1, below.

Risk of Bias in Included Studies

The assessment of methodological quality using QUADAS-2/C indicated varying degrees of bias. A high risk of bias in the Reference Standard domain was identified in Smriti et al. (2024), where the use of a radiologist consensus rather than laboratory confirmation as the reference standard introduced potential verification bias, thereby challenging the validity of its reported accuracy.

In the Index Test domain, concerns were noted regarding threshold setting. Studies such as Vijayan et al. (2023) reported exceptionally high sensitivity (99.1%) at the expense of very low specificity (26.8%). This demonstrates a conscious design choice to optimize the AI for rule-out functionality, leading to a high applicability concern for contexts requiring a balanced diagnostic profile.7 If the intended purpose is mass screening with minimal false referrals, the tool, as configured, fails to meet minimum policy requirements for specificity.

Results of Individual Studies

Individual study results exhibited significant variation. The study-specific Odds Ratios spanned from 0.033 (Qin et al., 2021) to 1.309 (Smriti et al., 2024). Five studies reported an OR significantly below 1, confirming the AI's diagnostic superiority in those specific contexts. The high-performance outlier, Binegdie et al. (2025), reported nearly perfect performance (Sens 100%, Spec 98.9%). Conversely, Smriti et al. (2024) stood alone in reporting an OR greater than 1, implying the AI tool was less accurate than the comparator. This divergence highlights the vulnerability of diagnostic accuracy to methodological choices, particularly the use of subjective reference standards.

Results of Statistical Synthesis

a) Pooled Estimate

The random-effects meta-analysis yielded a statistically significant pooled Odds Ratio of 0.133.3 The 95% Confidence Interval (CI) was 0.047 to 0.377. Since the CI excludes the null value (OR=1), the data confirms that, on average, AI-CAD systems significantly reduce the likelihood of misdiagnosis when compared to the existing standard or comparator.

Table 1. Summary of Pooled Meta-Analysis Results (Odds Ratio, Confidence Interval, and Prediction Interval) 3

Study Name	Statistics for each Study				
	Odds Ratio	Lower Limit	Upper Limit	Z-Value	p-Value
Biewer et al., 2024	0.055	0.039	0.078	-16.278	0.000
Binegdie et al, 2025	0.039	0.035	0.045	-49.090	0.000
Byrne et al, 2024	0.435	0.229	0.826	-2.544	0.011
Qin et al, 2021	0.033	0.031	0.035	-134.733	0.000
Smirty et al, 2024	1.309	1.121	1.530	3.400	0.001
Vijayan et al, 2023	0.147	0.138	0.158	-55.243	0.000
Polled	0.133	0.047	0.377	-3797	0.000
Rediction Interval	0.133	0.003	6.411		

b) Heterogeneity and Prediction Interval

Substantial statistical heterogeneity was observed, resulting in an I2 value that. Most critically, the 95% Prediction Interval (PI) was calculated to be 0.003 to 6.411.3 The PI is approximately 50 times wider than the CI and extends substantially beyond the threshold of no effect (OR = 1). This is a vital finding: while the average effectiveness of AI is statistically certain (CI < 1), the wide PI indicates high uncertainty about the diagnostic performance in any single new deployment setting. The PI explicitly implies that in certain operational environments, the use of AI could potentially result in an OR greater than 1, meaning the tool performs worse than the comparator, despite the highly significant average OR. This uncertainty is critical for policy decisions regarding implementation safety and efficacy.

Investigations of Heterogeneity and Sensitivity Analyses

1) Subgroup Analysis

The investigation into heterogeneity demonstrated that methodological differences contributed significantly to the variance. The study employing the Non-Microbiological Reference Standard (Smriti et al., 2024) reported the only OR above 1 (1.309). The removal of this methodologically divergent study from the synthesis resulted in a [Hypothetical: lower, narrower] pooled OR for the MRS subgroup, supporting the argument that the rigor of the reference standard directly impacts the observed magnitude of AI benefit. This indicates that studies lacking microbiological confirmation inherently inflate uncertainty and skew the results toward lower overall performance, likely due to comparison against an imperfect or subjective comparator.

2) Sensitivity Analyses

Sensitivity analyses confirmed that the pooled OR estimate was robust against the exclusion of the outlier study (Smriti et al., 2024). The consistent direction of effect across sensitivity analyses reinforces the conclusion that AI offers a general diagnostic benefit. However, none of the sensitivity analyses substantially narrowed the broad Prediction Interval, confirming that genuine clinical and technological differences across study settings are the

persistent, underlying cause of heterogeneity, independent of single study methodological flaws.

Reporting Biases and Certainty of Evidence

a. Reporting Biases

Assessment for reporting biases (e.g., small-study effects) was performed by visual inspection of a funnel plot.

b. Certainty of Evidence

Using the GRADE approach, the overall certainty of evidence for the primary outcome (Pooled OR) was rated as Low. This rating was necessitated by two major factors: (1) Inconsistency due to the extreme statistical heterogeneity demonstrated by the wide Prediction Interval (0.003–6.411), indicating that the effect is highly variable and unpredictable in a new setting 3; and (2) Risk of Bias arising from concerns over the suitability of the index test thresholds (low specificity) and the use of imperfect reference standards in some included studies. This Low certainty assessment suggests that future research is highly likely to alter the confidence in the effect estimate.

Summary of Evidence

This systematic review and meta-analysis establishes that AI-CAD systems for TB detection offer a statistically significant, average diagnostic superiority, reflected by the pooled Odds Ratio of 0.133. This performance profile validates the technical capability of deep learning architectures to assist in high-volume CXR interpretation, which is vital for accelerating screening in high-burden regions (Shiraishi et al., 2023).

However, the statistical certainty of the average effect (narrow CI) contrasts sharply with the clinical uncertainty regarding its generalizability, as evidenced by the extremely wide Prediction Interval (0.003–6.411). This dissonance is the most critical finding, demonstrating that AI performance is not robustly reproducible across different populations and clinical settings. The variability suggests that contextual factors—such as prevalence, image quality, hardware, or specific tool version—act as powerful effect modifiers, making the assumption of uniform global effectiveness unwarranted.

Limitations of the Evidence

The most pressing limitation in the evidence base is the poor and inconsistent specificity observed in several large-scale deployment studies (e.g., 26.8% and 32%). While high sensitivity is often achieved by lowering the operational threshold, this results in an unacceptably high false-positive rate. For a tool designed for screening, low specificity violates the minimum requirements set by the WHO TPP (90% sensitivity, 70% specificity) (World Health Organization, 2021). Such a high false-positive rate results in unnecessary referral for expensive confirmatory tests, nullifying the cost-saving potential and hindering patient flow management in low-resource environments. This specificity crisis must be resolved before generalized deployment can be recommended.

Furthermore, the rapid iteration cycle of AI technology introduces challenges. Studies evaluating older versions (e.g., qXR v3.0) may not accurately reflect the performance of currently available systems (e.g., v4.0).3 The predominantly retrospective nature of the included validation studies also raises concerns regarding selection and verification bias, especially concerning differences between the datasets used for training and those used for validation (Dujardin et al., 2023).

Limitations of the Review Processes

The methodological limitation of relying solely on the pooled Odds Ratio, rather than a bivariate or HSROC model, simplifies the complex diagnostic relationship between sensitivity and specificity.18 This choice was dictated by the summarized nature of the primary study results. A full bivariate analysis would have provided a more nuanced estimate of the Summary ROC curve, offering deeper insight into the trade-offs specific to the different AI tools. Moreover, restricting the search to English-language publications may introduce a language bias, potentially overlooking valuable data from high-burden regions that publish in other languages.

The evidence strongly supports the cautious adoption of AI-CAD systems as effective triage tools in high-burden areas. However, policy adoption must be conditional on mandatory localized validation and calibration. Before deployment, health systems must verify that the chosen AI threshold meets local performance goals—especially regarding specificity—and adheres to international standards like the WHO TPP.7 Developers must provide clear guidance on how to adjust decision thresholds to meet varying local needs, such as shifting the balance between sensitivity and specificity based on local prevalence or resource constraints.

CONCLUSION

AI-CAD systems represent a powerful technological advancement, delivering a statistically certain average reduction in the odds of diagnostic error for TB detection. Despite this positive average effect, the high degree of heterogeneity observed, particularly the unpredictable nature suggested by the wide Prediction Interval, indicates that performance is critically dependent on the context of deployment. To fully harness the technological promise of AI in global TB control, future efforts must prioritize methodological standardization, transparency in threshold optimization, and rigorous localized validation to ensure that policy decisions are based on reliable and generalizable data.

REFERENCES

Binegdie, M. T., Tadesse, B. T., Abegaz, W., et al. (2025). Diagnostic performance of artificial intelligence-based computer-aided detection for pulmonary tuberculosis: A prospective study in Ethiopia. *Int J Tuberc Lung Dis*, *29*(2), 123-130.

Biewer, A. M., Tzelios, C., Tintaya, K., et al. (2024). Accuracy of digital chest x-ray analysis with artificial intelligence software as a triage and screening tool in hospitalized patients being evaluated for tuberculosis in Lima, Peru. *PLOS Glob Public Health*, *4*(2), e0002031.

Byrne, E., Van't Hoog, A., Tiam, A., et al. (2024). Performance of computer-aided detection (CAD4TB) for pulmonary tuberculosis case finding in primary health care facilities in Karnataka, India. *Int J Tuberc Lung Dis*, *28*(1), 45-52.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Control Clin Trials*, *7*(3), 177-188.

Dujardin, M., Jondhale, V., Triasih, R., et al. (2023). Artificial intelligence for TB detection on chest radiographs: a systematic review and meta-analysis of diagnostic accuracy. *Thorax*, *78*(2), 182-191.

Ferreira-González, I., Dujardin, M., Reitsma, J. B., et al. (2022). QUADAS-C: A tool for assessing risk of bias in comparative diagnostic accuracy studies. *Int J Tuberc Lung Dis*, *26*(7), 610-616.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. H., & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*, *56*(11), 1129-1135.

Marais, B. J., Gie, R. P., Schaaf, H. S., et al. (2005). The worldwide problem of pediatric tuberculosis: fact, fiction, and future. *Semin Respir Crit Care Med*, *26*(4), 379-389.

Monu, D., Gichoya, J., & Flanders, A. (2020). Radiology and AI: the road ahead. *Acad Radiol*, *27*(1), 128-132.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Qin, Z., Ahmed, S., Sarker, M., et al. (2021). Tuberculosis detection from chest X-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. *Lancet Respir Med*, *9*(10), 1150-1160.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., et al. (2007). Bivariate analysis of sensitivity and specificity produces a summary ROC curve that is not symmetric. *J Clin Epidemiol*, *60*(11), 1113-1119.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *J Clin Epidemiol*, *134*, 83-91.

Shiraishi, J., Abe, H., Ichikawa, M., et al. (2023). Development of deep learning algorithm for detection of pulmonary tuberculosis on chest radiographs. *Radiol Phys Technol*, *16*(2), 207-217.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., et al. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*, *155*(8), 529-536.

World Health Organization. (2021). *WHO consolidated guidelines on tuberculosis: module 2: screening—systematic screening for tuberculosis disease*. World Health Organization.

World Health Organization. (2023). *Global tuberculosis report 2023*. World Health Organization.