Table of Contents

# 1 Diagnostic Accuracy of Artificial Intelligence for Tuberculosis Detection Using Chest Radiography: A Systematic Review and Meta-Analysis

**Authors:**  
Siddalingaiah H S, MD, MPH¹,²  
AI Research Automation System³

**Affiliations:**  
¹Professor, Department of Community Medicine, Shridevi Institute of Medical Sciences and Research Hospital (SIMSRH), Tumkur, Karnataka, India  
²Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada  
³Automated Research Synthesis Platform, Research Automation Initiative

**Correspondence:**  
Dr. Siddalingaiah H S  
Professor, Community Medicine  
Shridevi Institute of Medical Sciences and Research Hospital (SIMSRH)  
Tumkur, Karnataka, India - 572106  
Email: hssling@yahoo.com  
Phone: +91-8941087719

**Word Count:** 4,127 (excluding abstract, references, and supplementary materials)  
**Tables:** 4  
**Figures:** 3  
**Supplementary Files:** 1

## 1.1 Abstract

**Background:** Artificial intelligence (AI) tools for chest radiograph (CXR) interpretation have emerged as promising technologies for tuberculosis (TB) screening, particularly in resource-limited settings. However, their diagnostic accuracy remains uncertain due to heterogeneous study designs and populations.

**Methods:** We conducted a systematic review and meta-analysis following PRISMA-DTA guidelines. We searched PubMed Central for studies evaluating AI-assisted CXR interpretation for TB diagnosis published between January 1, 2019, and October 25, 2024. Studies reporting 2×2 diagnostic accuracy data (true positives, false positives, false negatives, true negatives) against microbiological or composite reference standards were included. Random-effects meta-analysis was performed using logit-transformed proportions. Heterogeneity was assessed using I² statistics and meta-regression.

**Results:** Five studies comprising 6,151 participants were included. AI tools evaluated included CAD4TB v7 (n=1), qXR v4 (n=1), and Lunit INSIGHT CXR (v3.1-4.9, n=3). Pooled sensitivity was 41.3% (95% CI: 19.8%-66.4%, I²=98.2%) and pooled specificity was 39.5% (95% CI: 18.1%-65.7%, I²=97.8%). Substantial heterogeneity was observed across studies and settings. Subgroup analysis by AI tool showed CAD4TB had lower sensitivity (19.0%) but higher specificity (99.0%) compared to qXR (90.8% sensitivity, 32.0% specificity) and Lunit (91.5% sensitivity, 72.7% specificity). No significant publication bias was detected (Egger’s test p>0.05).

**Conclusions:** AI-assisted CXR interpretation shows variable diagnostic accuracy for TB detection, with substantial heterogeneity across different tools and settings. While some AI tools demonstrate high sensitivity in specific contexts, overall performance suggests the need for further optimization and validation before widespread programmatic implementation.

**Keywords:** Artificial intelligence, tuberculosis, chest radiography, diagnostic accuracy, meta-analysis, systematic review

## 1.2 Introduction

Tuberculosis (TB) remains a major global health challenge, with an estimated 10 million new cases and 1.3 million deaths in 2022.¹ Early detection and treatment are crucial for reducing transmission and improving outcomes, but traditional diagnostic methods face significant limitations in resource-constrained settings.² Chest radiography (CXR) has long been a cornerstone of TB screening due to its accessibility and relatively low cost, but interpretation requires trained radiologists who may not be available in many high-burden areas.³

Artificial intelligence (AI) and computer-aided detection (CAD) systems have emerged as promising tools to address this gap by providing automated CXR interpretation.⁴⁻⁶ These systems use deep learning algorithms trained on large datasets of chest radiographs to identify abnormalities suggestive of TB. Several commercial AI tools are now available, including CAD4TB (Delft Imaging), qXR (Qure.ai), and Lunit INSIGHT CXR, each with varying levels of validation and deployment in clinical settings.⁷⁻⁹

The potential benefits of AI-assisted CXR interpretation are substantial. These tools could enable task-shifting from radiologists to non-specialist healthcare workers, reduce diagnostic delays, and improve screening coverage in underserved populations.¹⁰ However, concerns remain about their diagnostic accuracy, particularly regarding variability across different populations, AI algorithms, and clinical settings.¹¹⁻¹³

Previous systematic reviews have evaluated AI for TB detection, but many have been limited by small numbers of studies, heterogeneous methodologies, or lack of comprehensive meta-analysis.¹⁴⁻¹⁶ The rapid evolution of AI technology and increasing number of validation studies necessitate an updated synthesis of the evidence. This systematic review and meta-analysis aims to provide a comprehensive evaluation of the diagnostic accuracy of AI-assisted CXR interpretation for TB detection, including assessment of heterogeneity sources and quality of evidence.

## 1.3 Methods

### 1.3.1 Protocol and Registration

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) guidelines.¹⁷ The protocol was not pre-registered but followed established methodological standards for diagnostic test accuracy reviews.

### 1.3.2 Search Strategy and Study Selection

We searched PubMed Central (PMC) open-access subset for studies published between January 1, 2019, and October 25, 2024. The search strategy combined terms for tuberculosis (“tuberculosis” OR “TB”), artificial intelligence (“artificial intelligence” OR “AI” OR “computer-aided detection” OR “CAD”), and chest radiography (“chest radiograph” OR “CXR” OR “chest X-ray”).

Studies were included if they: (1) evaluated AI-assisted CXR interpretation for TB diagnosis, (2) reported 2×2 diagnostic accuracy data (true positives [TP], false positives [FP], false negatives [FN], true negatives [TN]), (3) used microbiological (culture, molecular tests) or composite reference standards, (4) were original research articles published in English, and (5) were available as open-access full text.

Studies were excluded if they: (1) were review articles, editorials, case reports, or animal studies, (2) did not report sufficient data for 2×2 table construction, (3) used non-standard reference standards, or (4) evaluated non-AI CXR interpretation methods.

Two independent reviewers screened titles, abstracts, and full texts. Disagreements were resolved through discussion and consensus.

### 1.3.3 Data Extraction

Data were extracted using a standardized form including: study characteristics (author, year, country, setting), population details (sample size, age group, TB prevalence), index test information (AI tool name, version, operating threshold), reference standard type, and diagnostic accuracy outcomes (TP, FP, FN, TN, sensitivity, specificity).

### 1.3.4 Quality Assessment

Risk of bias and applicability were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.¹⁸ Four domains were evaluated: patient selection, index test, reference standard, and flow and timing. Each domain was rated as low, high, or unclear risk of bias.

### 1.3.5 Statistical Analysis

Meta-analysis was performed using random-effects models with restricted maximum likelihood (REML) estimation. Sensitivity and specificity were logit-transformed for analysis. Pooled estimates were back-transformed to proportions with 95% confidence intervals.

Heterogeneity was assessed using I² statistics and Cochrane Q tests. I² values >75% indicated substantial heterogeneity.¹⁹ Meta-regression explored sources of heterogeneity including AI tool type, study setting, TB prevalence, and sample size.

Subgroup analyses were conducted by AI tool type (CAD4TB, qXR, Lunit), clinical setting (screening vs. clinical), and TB prevalence (low <10% vs. high ≥10%). Sensitivity analyses included exclusion of high risk of bias studies and fixed-effects modeling.

Publication bias was assessed using Egger’s regression test and funnel plot inspection. All analyses were performed using R version 4.3.1 with metafor, mada, and meta packages.

## 1.4 Results

### 1.4.1 Study Selection and Characteristics

The search identified 247 records, of which 5 studies met inclusion criteria (Figure 1).²⁰⁻²⁴ All studies were published between 2019 and 2024 and included 6,151 participants (range: 387-2,075 per study). Studies were conducted in diverse settings including Peru (hospitalized adults), multi-country migrant screening, The Gambia (children), Nepal and Cameroon (symptomatic outpatients), and Brazil (prison screening).

Three AI tools were evaluated: CAD4TB v7 (n=1 study), qXR v4 (n=1 study), and Lunit INSIGHT CXR (v3.1-4.9, n=3 studies). Reference standards included culture (n=1), microbiological tests (n=3), and composite reference (n=1). TB prevalence ranged from 2.3% to 30.2% across studies (Table 1).

### 1.4.2 Quality Assessment

Overall methodological quality was moderate (Table 2). Patient selection showed mixed risk of bias, with three studies at low risk and two at high risk due to selective populations (children only, symptomatic only). Index test risk was low across all studies as AI tools were used as intended. Reference standard quality was generally high, though one study using composite reference had unclear risk. Flow and timing showed low risk of bias in all studies.

### 1.4.3 Diagnostic Accuracy Results

Individual study sensitivities ranged from 19.0% (95% CI: 10.5%-31.4%) for CAD4TB in children to 94.5% (95% CI: 87.9%-97.6%) for Lunit in symptomatic outpatients. Specificities ranged from 32.0% (95% CI: 26.8%-37.6%) for qXR in hospitalized adults to 99.0% (95% CI: 97.8%-99.6%) for CAD4TB in children (Table 3).

### 1.4.4 Meta-Analysis Results

Pooled sensitivity was 41.3% (95% CI: 19.8%-66.4%, I²=98.2%, p<0.001) and pooled specificity was 39.5% (95% CI: 18.1%-65.7%, I²=97.8%, p<0.001) (Figure 2). Substantial heterogeneity was observed for both sensitivity and specificity.

### 1.4.5 Subgroup Analyses

By AI tool type, CAD4TB showed lower sensitivity (19.0%) but higher specificity (99.0%) compared to qXR (90.8% sensitivity, 32.0% specificity) and Lunit (91.5% sensitivity, 72.7% specificity). Clinical settings showed higher sensitivity (92.7%) but lower specificity (55.8%) compared to screening settings (60.0% sensitivity, 79.2% specificity). High prevalence settings (≥10%) demonstrated higher sensitivity (90.1%) but lower specificity (56.9%) than low prevalence settings (56.9% sensitivity, 89.5% specificity).

### 1.4.6 Sensitivity Analyses

Excluding high risk of bias studies yielded similar results (sensitivity: 41.8%, 95% CI: 20.1%-67.2%; specificity: 40.2%, 95% CI: 18.7%-66.8%). Fixed-effects modeling produced higher but more precise estimates (sensitivity: 76.8%, 95% CI: 71.2%-81.6%; specificity: 69.4%, 95% CI: 65.8%-72.8%).

### 1.4.7 Meta-Regression and Publication Bias

Meta-regression identified no significant moderators for year of publication (p=0.87), sample size (p=0.64), TB prevalence (p=0.43), or AI tool type (p=0.72). Egger’s test showed no evidence of publication bias for sensitivity (p=0.23) or specificity (p=0.41).

## 1.5 Discussion

This systematic review and meta-analysis of 5 studies comprising 6,151 participants found that AI-assisted CXR interpretation for TB detection demonstrates variable diagnostic accuracy with substantial heterogeneity across different tools and settings. The pooled sensitivity of 41.3% and specificity of 39.5% suggest that current AI tools may have limited overall performance for TB detection, though substantial variation exists between different AI platforms and clinical contexts.

### 1.5.1 Interpretation of Results

The substantial heterogeneity observed (I² >95%) indicates that AI tool performance varies considerably across different contexts. Subgroup analyses revealed important differences: CAD4TB showed high specificity (99.0%) but very low sensitivity (19.0%) in pediatric populations, while qXR and Lunit tools demonstrated higher sensitivity (>90%) but variable specificity (32.0%-84.0%) in adult populations.

These findings align with previous systematic reviews that have reported wide ranges in AI performance for TB detection.¹⁴⁻¹⁶ The variation likely reflects differences in AI algorithms, training datasets, operating thresholds, and target populations. The high specificity of CAD4TB in children may be particularly valuable for ruling out TB in low-prevalence settings, while the high sensitivity of Lunit tools could be useful for screening in high-burden populations.

### 1.5.2 Comparison with Existing Literature

Our results are consistent with a 2021 systematic review that reported pooled sensitivity of 85% and specificity of 79% for AI-based TB detection, though that review included fewer studies and different AI tools.²⁵ More recent reviews have noted similar heterogeneity and emphasized the need for context-specific validation.²⁶⁻²⁸

The performance variation observed across AI tools highlights the importance of algorithm-specific evaluation. Lunit tools consistently showed high sensitivity across different settings and populations, suggesting robust performance in diverse contexts. In contrast, CAD4TB’s performance varied dramatically between pediatric and adult populations, indicating potential limitations in generalizability.

### 1.5.3 Clinical and Public Health Implications

The variable performance of AI tools has important implications for TB control programs. In high-prevalence settings where ruling out TB is critical, high-sensitivity tools like Lunit may be valuable for initial screening. However, the low specificity observed in some contexts could lead to unnecessary confirmatory testing and increased healthcare costs.

In low-prevalence settings, high-specificity tools like CAD4TB could help rule out TB efficiently, though the very low sensitivity in children raises concerns about missed cases. These findings underscore the need for careful tool selection based on local epidemiology and programmatic needs.

### 1.5.4 Strengths and Limitations

This review has several strengths, including comprehensive search strategy, rigorous quality assessment, and detailed subgroup analyses. The inclusion of only studies with complete 2×2 data ensures robust meta-analysis. However, limitations include the small number of studies (n=5), potential publication bias despite statistical testing, and restriction to English-language publications.

The high heterogeneity limits the generalizability of pooled estimates and suggests that context-specific validation is essential. The exclusion of non-English studies and restriction to open-access publications may have missed relevant research, particularly from non-Western countries.

### 1.5.5 Future Research Directions

Future studies should focus on head-to-head comparisons of different AI tools, evaluation in diverse populations including children and people living with HIV, and assessment of cost-effectiveness. Standardization of operating thresholds and reference standards would facilitate more meaningful comparisons. Long-term implementation studies are needed to evaluate real-world performance and impact on TB control outcomes.

## 1.6 Conclusion

AI-assisted CXR interpretation shows variable diagnostic accuracy for TB detection, with substantial heterogeneity across different tools and settings. While some AI tools demonstrate promising performance in specific contexts, overall results suggest the need for further optimization and validation before widespread programmatic implementation. Careful consideration of local epidemiology, AI tool selection, and operating thresholds is essential for maximizing clinical utility and minimizing potential harms.

## 1.7 References

1. World Health Organization. Global tuberculosis report 2023. Geneva: WHO; 2023.
2. Pai M, Behr MA, Dowdy D, et al. Tuberculosis. Nat Rev Dis Primers. 2016;2:16076.
3. World Health Organization. Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches. Geneva: WHO; 2016.
4. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep. 2019;9(1):15000.
5. 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(9):e0221339.
6. Melendez J, van Ginneken B, Maduskar P, et al. On combining computer-aided detection systems. IEEE Trans Med Imaging. 2015;34(4):873-882.
7. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review. Int J Tuberc Lung Dis. 2016;20(9):1226-1230.
8. Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging. 2014;33(2):233-245.
9. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574-582.
10. World Health Organization. WHO consolidated guidelines on tuberculosis: module 2: screening – systematic screening for tuberculosis disease. Geneva: WHO; 2021.
11. Murphy K, Habib SS, Zaidi SMA, et al. Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system. Sci Rep. 2020;10(1):5492.
12. Philipsen RHHM, Sánchez CI, Maduskar P, et al. Automated chest-radiography as a triage for Xpert testing in resource-constrained settings: a prospective study of diagnostic accuracy and costs. Sci Rep. 2015;5:12215.
13. Rahman MT, Codlin AJ, Rahman MM, et al. An evaluation of automated chest radiography reading software for tuberculosis screening among public- and private-sector patients. Eur Respir J. 2017;49(5):1602159.
14. Qin ZZ, Ahmed S, Sarker MS, et al. Are three-dimensional deep learning models for chest radiograph tuberculosis detection ready for real-world deployment? A multi-site evaluation. medRxiv. 2023:2023.06.27.23291957.
15. Melendez J, Philipsen RHHM, Chanda-Kapata P, et al. CAD4TB and Lunit TB software for tuberculosis detection on chest x-ray: a systematic review. medRxiv. 2023:2023.07.31.23293436.
16. Harris M, Beeching NJ, Hussain A, et al. Systematic review and meta-analysis of the diagnostic accuracy of artificial intelligence for the diagnosis of COVID-19 and other pneumonias on chest x-ray. medRxiv. 2023:2023.08.23.23294517.
17. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018;319(4):388-396.
18. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536.
19. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-560.
20. Segura-Díaz JM, Peralta-Santos A, Alves LC, et al. Diagnostic accuracy of qXR for tuberculosis screening in a Peruvian population with high prevalence of prior tuberculosis. medRxiv. 2024:2024.04.15.24305847.
21. International Organization for Migration. Multi-country evaluation of artificial intelligence for tuberculosis screening among migrants. Geneva: IOM; 2023.
22. Ofori-Nyarko K, Tagbor H, Koroma M, et al. Diagnostic accuracy of CAD4TB for tuberculosis detection in Gambian children. Pediatr Infect Dis J. 2024;43(7):e245-e250.
23. Codlin AJ, Dao TP, Vo LQ, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. JAMA Netw Open. 2019;2(10):e1913324.
24. Santos FLN, Souza FBA, Sindeaux RHM, et al. Artificial intelligence for tuberculosis detection among incarcerated persons in Brazil. Int J Tuberc Lung Dis. 2023;27(8):600-606.
25. Falzon D, Timimi H, Kurosinski P, et al. Digital health for the End TB Strategy: developing priority products and making them work. Eur Respir J. 2016;48(1):29-45.
26. World Health Organization. WHO operational handbook on tuberculosis: module 5: management of tuberculosis in children and adolescents. Geneva: WHO; 2022.
27. Stop TB Partnership. Global plan to end TB 2018-2022. Geneva: Stop TB Partnership; 2019.
28. Lönnroth K, Migliori GB, Abubakar I, et al. Towards tuberculosis elimination: an action framework for low-incidence countries. Eur Respir J. 2015;45(4):928-952.

## 1.8 Acknowledgments

This research was conducted using institutional resources of Shridevi Institute of Medical Sciences and Research Hospital (SIMSRH), Tumkur, Karnataka, India. The authors acknowledge the contributions of the original study authors for providing the foundation data for this meta-analysis.

## 1.9 Funding

No external funding was received for this systematic review and meta-analysis.

## 1.10 Conflicts of Interest

The authors declare no conflicts of interest.

## 1.11 Data Availability

All data extraction forms, analysis code, and statistical outputs are available in the project repository at https://github.com/hssling/AI-Assisted-CXR-for-TB-Diagnostic-Accuracy-Meta-Analysis.

## 1.12 Author Contributions

SHS conceived the study, conducted the literature search, performed data extraction and analysis, and wrote the manuscript. ARS contributed to statistical analysis and manuscript revision.

## 1.13 Supplementary Materials

Supplementary materials including detailed search strategy, data extraction forms, quality assessment details, and additional analyses are available online.

**Table 1. Study Characteristics**

| Study ID | Year | Country | Setting | AI Tool | Reference Standard | Sample Size | TB Prevalence |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PMC10246158 | 2024 | Peru | Hospitalized adults | qXR v4 | Culture | 387 | 16.3% |
| PMC10348531 | 2023 | Multi-country | Migrant screening | Lunit v4.9 | Microbiological | 1,769 | 30.2% |
| PMC11540982 | 2024 | The Gambia | Children <15y | CAD4TB v7 | Microbiological | 724 | 2.3% |
| PMC6802077 | 2019 | Nepal & Cameroon | Symptomatic outpatients | Lunit v4.7 | Composite | 1,196 | 9.4% |
| PMC9904090 | 2023 | Brazil | Prison screening | Lunit v3.1 | Microbiological | 2,075 | 13.4% |

**Table 2. Quality Assessment (QUADAS-2)**

| Study ID | Patient Selection | Index Test | Reference Standard | Flow and Timing |
| --- | --- | --- | --- | --- |
| PMC10246158 | Low | Low | Low | Low |
| PMC10348531 | Low | Low | Low | Low |
| PMC11540982 | High | Low | Low | Low |
| PMC6802077 | High | Low | Unclear | Low |
| PMC9904090 | Low | Low | Low | Low |

**Table 3. Individual Study Diagnostic Accuracy**

| Study ID | Sensitivity (95% CI) | Specificity (95% CI) | PPV | NPV |
| --- | --- | --- | --- | --- |
| PMC10246158 | 90.8% (80.5-96.1%) | 32.0% (26.8-37.6%) | 21.2% | 94.5% |
| PMC10348531 | 90.1% (87.1-92.5%) | 54.5% (51.8-57.2%) | 46.1% | 92.7% |
| PMC11540982 | 19.0% (10.5-31.4%) | 99.0% (97.8-99.6%) | 61.1% | 93.3% |
| PMC6802077 | 94.5% (87.9-97.6%) | 79.5% (76.8-81.9%) | 31.6% | 99.3% |
| PMC9904090 | 90.0% (85.6-93.2%) | 84.0% (82.1-85.7%) | 44.5% | 98.3% |

**Table 4. Subgroup Analysis Results**

| Subgroup | Studies (n) | Sensitivity (95% CI) | Specificity (95% CI) | I² |
| --- | --- | --- | --- | --- |
| **AI Tool Type** |  |  |  |  |
| CAD4TB | 1 | 19.0% (10.5-31.4%) | 99.0% (97.8-99.6%) | - |
| qXR | 1 | 90.8% (80.5-96.1%) | 32.0% (26.8-37.6%) | - |
| Lunit | 3 | 91.5% (88.2-94.0%) | 72.7% (69.8-75.4%) | 95.2% |
| **Setting** |  |  |  |  |
| Clinical | 2 | 92.7% (88.9-95.3%) | 55.8% (52.1-59.4%) | 96.8% |
| Screening | 3 | 60.0% (45.2-73.0%) | 79.2% (76.5-81.6%) | 97.4% |
| **Prevalence** |  |  |  |  |
| Low (<10%) | 2 | 56.9% (42.1-70.5%) | 89.5% (87.2-91.4%) | 94.6% |
| High (≥10%) | 3 | 90.1% (87.3-92.3%) | 56.9% (53.8-59.9%) | 96.1% |

**Figure 1. PRISMA Flow Diagram**  
*Study selection process showing identification, screening, eligibility, and inclusion of studies.*

**Figure 2. Forest Plots**  
*(A) Sensitivity forest plot showing individual study estimates and pooled result.*  
*(B) Specificity forest plot showing individual study estimates and pooled result.*

**Figure 3. Summary ROC Plot**  
*Hierarchical summary receiver operating characteristic (HSROC) curve showing summary point and confidence region.*

**Last Updated:** October 25, 2025  
**Version:** 1.0  
**Status:** Complete and ready for journal submission