INT J TUBERC LUNG DIS 20(9):1226–1230

http://dx.doi.org/10.5588/ijtld.15.0926

# Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review

**T. Pande,* C. Cohen,* M. Pai,* F. Ahmad Khan*†**

***McGill International TB Center, Montreal, †Respiratory Epidemiology and Clinical Research Unit, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada**

SUMMARY

**OBJECTIVE: To systematically review the diagnostic accuracy of computer-aided detection (CAD) of pulmonary tuberculosis (PTB) on digital chest radiographs (CXR).**

**DESIGN: We searched four databases for articles published between January 2010 and December 2015 comparing CAD of PTB on CXR to a microbiologic reference standard (smear, culture or polymerase chain reaction). We collected and summarised data on study design, CAD software and diagnostic accuracy (sensitivity, specificity, area under the curve [AUC]).**

**RESULTS: We included 5 of 455 articles identified by searching databases. PTB prevalence ranged from 18% to 60%, and human immunodeficiency virus (HIV) prevalence from 33% to 68%. All articles evaluated CAD4TB, the only commercially available software. AUC ranged from 0.71 to 0.84. Software settings that increased sensitivity resulted in important reductions in specificity, and vice versa. Risk of bias was low in prospective studies ($n = 2$), and high in retrospective studies ($n = 3$).**

**CONCLUSION: Evidence assessing CAD's diagnostic accuracy is limited by the small number of studies, most of which have important methodological limitations, the availability and evaluation of only one software programme, and limited generalisability to settings where PTB and HIV are less prevalent. Additional research is required.**

**KEY WORDS**: digital X-rays; chest X-ray; computer-aided detection; diagnostic technology; TB diagnostics

AS CHEST RADIOGRAPHY (CXR) is a highly sensitive and moderately specific test for active pulmonary tuberculosis (PTB), it is useful as a screening tool.[1] Challenges to the use of CXR in resource-limited settings include high-cost equipment and a paucity of professionals to interpret images.[2,3] These barriers could be overcome by using digital CXR and software programmes that can detect radiographic abnormalities compatible with PTB, termed 'computer-aided detection', or CAD. In settings where CXR is already incorporated into TB diagnostic algorithms, use of CAD systems could eliminate problems of poor inter-reader reliability,[4] and potential delays in interpretation if relying on human readers. We conducted a systematic review of studies evaluating the diagnostic accuracy of CAD software for microbiologically confirmed PTB.

## METHODS

We searched PubMed, EMBASE, SCOPUS and Engineering Village, using a sensitive search strategy formulated in consultation with a medical librarian (see Appendix: Search Strategy).[1] Search terms were kept as broad as possible, and no language restrictions were applied. As CAD is a nascent technology, the search period was limited to papers published after 1 January 2010, and included articles published up to 31 December 2015.

Studies were included if they estimated the diagnostic accuracy of CAD software for the detection of PTB compared to a microbiological reference standard (smear, culture or Xpert® MTB/RIF [Cepheid, Sunnyvale, CA, USA]). Studies were excluded if they reported CAD for diagnostic modalities other than CXR, or if they did not use a microbiological reference standard to determine CAD accuracy.

Two reviewers independently extracted data using a standardised form. We collected information on the year of publication; geographic location; source of

TP and CC contributed equally to this work

[1] The appendix is available in the online version of this article, at http://www.ingentaconnect.com/content/iuatld/ijtld/2016/00000020/00000009/00019

Correspondence to: Faiz Ahmad Khan, Respiratory Epidemiology and Clinical Research Unit, Research Institute of the McGill University Health Centre, Office 413, 4th Floor, 2155 Guy Street, Montréal, Québec H3H 2R9, Canada. Fax: (+1) 514 843 2083. e-mail: faiz.ahmadkhan@mcgill.ca

*Article submitted 17 November 2015. Final version accepted 8 May 2016.*

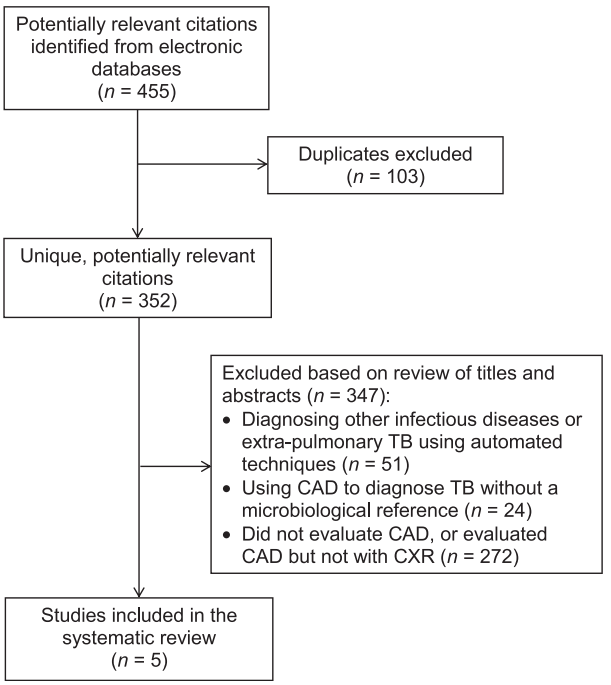


**Figure** Study selection flow chart. TB = tuberculosis; CAD = computer-aided detection; CXR = chest X-ray.

financial support; study design (e.g., retrospective vs. prospective vs. case-control); inclusion/exclusion criteria; method of generating digital radiographs; name, version and method of scoring used by the CAD software program; tests used as the microbiological reference; characteristics of the study populations; and the following measures of diagnostic accuracy: sensitivity, specificity and area under the receiver operating characteristic curve (AUC). For studies that also compared CAD to human interpretation, we recorded the number and expertise of the human readers, and results of the comparison.

We evaluated each study for risk of bias and applicability concerns (Appendix Figure A) in the three QUADAS-2 domains: patient selection, performance of the index test and performance of the reference test.[5] Because of variability in methods and results, no meta-analysis was performed.

## RESULTS

### *Study selection*

As shown in the Figure, our search identified 455 relevant citations; after de-duplication 352 remained, and after review, 5 articles were included.

### *CAD software*

All five studies were co-authored by members of the Diagnostic Image Analysis Group at Radboud University Medical Center, Nijmegen, The Netherlands, and evaluated this group's proprietary software, CAD4TB.[6–10] CAD4TB was developed using machine learning methods and sets of 'training' radiographs. It quantifies various image characteristics of a CXR to compute an 'abnormality score' that can take integer values ranging from 0 to 100. Higher abnormality scores indicate greater likelihood of PTB. For binary classification, a 'threshold score' is selected—scores less than the threshold are interpreted as indicating that PTB has been ruled out, whereas scores greater than or equal to the threshold indicate a possible diagnosis of PTB. The CAD4TB programme was developed with training CXRs from a database of presumed TB patients from Zambia and South Africa,[8–10] and its diagnostic performance has been evaluated in CXRs of patients from Zambia[8,9] Tanzania,[10] England[6] and South Africa.[6,7]

### *Study design*

The diagnostic accuracy of CAD was evaluated prospectively in two studies,[7,9] and retrospectively in three studies.[6,8,10] Three studies used microbiological criteria—smear or culture[7,8] or Xpert[9]—as the reference standard. One study used both culture and clinical criteria as the reference standard but did not provide details on the latter.[6] One study used a positive culture result as the reference standard for confirming PTB, but required negative smear/culture plus no evidence of PTB during 5 months of follow-up to exclude PTB;[10] in this study, patients with negative smears and negative cultures who did not meet the follow-up criteria were excluded when calculating CAD4TB's diagnostic accuracy.

As shown in the Table, HIV infection was highly prevalent in four of the six study populations, and was not reported in the others. The pre-test probability of PTB ranged from 18.3% to 60%.

### *Diagnostic accuracy of CAD*

All studies reported the AUC of the abnormality score as a measure of accuracy (Table). The highest AUC was reported with the newer version of CAD4TB (v 3.07); however, the AUC may have been overestimated in the study, as a large proportion of enrolled participants was excluded from the analyses of diagnostic accuracy.[10]

Three studies also reported sensitivity and specificity using threshold scores (Table).[8–10] The abnormality score selected as the threshold, and the rationale for this choice, varied across studies. Threshold scores achieving a sensitivity of $\geqslant 0.90$ were associated with moderate to poor specificity. One study reported specificity of $\geqslant 0.50$ at five different threshold scores; however, specificity may have been overestimated in this study due to the exclusion of a large number of patients.[10]

### *CAD vs. human interpretation*

In studies where human readers scored radiographs on a scale of 0 to 100, the older version of CAD4TB

**Table** Characteristics and results of studies reporting the diagnostic accuracy of CAD for PTB

| | | | | | | | | Diagnostic accuracy of CAD for microbiologically confirmed PTB | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year of study* | CAD software, version | Country | Reference standard | Patients enrolled *n* | Sample size for diagnostic accuracy | HIV *n* (%) | PTB *n* (%) | AUC (95%CI) | Threshold score (scale of 0–100) | Rationale for selection of threshold score | Sensitivity % (95%CI) | Specificity % (95%CI) |
| 2011[8] | CAD4TB, 1.08 | Zambia | Smear and culture | 161[†] | 161 | 110 (68) | 97 (60) | 0.73 (0.64–0.80) | NR | Set to achieve the same specificity as the field officer (0.41) | 0.86 (0.75–0.94) | 0.41[‡] |
| 2013[9] | CAD4TB, 1.08 | Zambia | Xpert | 458 | 350 | 190 (54) | 96 (33) | 0.71 (0.66–0.77) | ⩾61 | Threshold score selected based on AUC generated in pilot study | 1.0 (0.96–1.0) | 0.23 (0.18–0.29) |
| NR[10] | CAD4TB v 3.07 | Tanzania | Culture | 894 | 427 | 177 (41) | 194 (23) | 0.84 (0.80–0.88) | ⩾23 | NR | 0.95 (0.91–0.98) | 0.33 (0.27–0.39) |
| | | | | | | | | | ⩾37 | NR | 0.91 (0.86–0.94) | 0.52 (0.46–0.59) |
| | | | | | | | | | ⩾56 | NR | 0.85 (0.79–0.90) | 0.69 (0.62–0.75) |
| | | | | | | | | | ⩾74 | NR | 0.77 (0.71–0.83) | 0.79 (0.74–0.84) |
| | | | | | | | | | ⩾89 | NR | 0.62 (0.55–0.69) | 0.85 (0.80–0.89) |
| | | | | | | | | | ⩾95 | NR | 0.47 (0.40–0.54) | 0.94 (0.91–0.97) |
| 2013[6] | CAD4TB, version not reported | England | Culture and clinical | NR | 200 | NR | 87 (43) | 0.868 (NR) | NR | Not applicable | NR | NR |
| | | South Africa | Culture | NR | 200 | NR | 66 (33) | 0.741 (NR) | NR | Not applicable | NR | NR |
| NR[7] | CAD4TB v 3.07 | South Africa | Culture | 419 | 388 | 128 (33) | 71 (18) | 0.79 (NR) | NR | Not applicable | NR | NR |

* Last calendar year in which participants were enrolled.
[†] Data collected for a WHO specimen bank.[11] Only Zambian patients' data were analysed, and only those in whom CXR could be 'linked with clinical information'. The number in whom CXR could not be linked was not reported.
[‡] Study sought to compare the CAD4TB software's performance to that of clinical officers and a field officer reading CXR. The investigators set the CAD4TB threshold to achieve the same specificity as the field officer (0.41).
CAD = computer-aided detection; PTB = pulmonary tuberculosis; HIV = human immunodeficiency virus; AUC = area under the receiver operating characteristic curve; CI = confidence interval; NR = not reported; WHO = World Health Organization; CXR = chest X-ray.

had an AUC similar to that of clinical officers,[8] and the newer version performed similarly to expert radiologists.[7] When the CAD4TB software's threshold score was set to achieve the same specificity as a non-expert clinician identifying any radiographic abnormality, the software achieved the same sensitivity as other non-expert clinicians.[8] When the threshold was set to achieve the same sensitivity as human readers identifying abnormalities consistent with or highly suggestive of active TB, CAD4TB's specificity was lower than that of a radiologist and greater than that of a clinical officer.[10]

### *Quality assessment*

Risk of bias was considered low for all three QUADAS-2 domains in the two prospective studies.[7,9] For the other studies, major concerns about risk of bias in patient selection were related to use of a case-control design[6] and to inappropriate exclusion of patients from analyses.[10] Concerns about risk of bias in the performance of the index test stemmed from the use of threshold scores that were not pre-specified, or if information on pre-specification was lacking;[8,10] for the reference standard, concerns were related to the use of combined clinical and microbiological criteria to define PTB status.[6,10] Applicability concerns were related to whether results could be generalised to populations other than those that had been used to train the software or to settings where HIV prevalence or pre-test probability of PTB are lower than in the patient populations where CAD software was evaluated.

## DISCUSSION

Advances in digital imaging and CAD software have the potential to improve case finding. However, the evidence base to support the use of CAD for PTB diagnosis is limited by the small number of published studies, most of which have methodological limitations, and all of which used the same software programme for CAD. In this limited evidence base, CAD4TB has been reported to achieve the same diagnostic accuracy as non-expert clinicians,[8] and the newer version has approached the accuracy of an expert.[7,10] The data also suggest that selecting a high sensitivity threshold results in a substantial fall in specificity, and vice versa.

CAD software capable of achieving both high sensitivity and specificity would mark a major leap forward for this new technology. The Radboud group and their collaborators at the University of Cape Town, South Africa, recently evaluated an algorithm that used CAD4TB radiograph analysis as a test for triaging patients with presumed PTB towards confirmatory testing with Xpert.[7] They reported that increasing the threshold score would have allowed for reductions in the number of presumed PTB patients undergoing confirmatory testing, but at the expense of a greater number of missed cases of PTB (e.g., 11% of PTB cases would have been missed using a threshold score that would have triaged 40% of suspects for Xpert testing). The algorithm missed fewer cases in non-HIV-infected patients, in part because Xpert—the downstream test—is more sensitive in this group. Similar studies that also account for the costs and clinical consequences of missed TB cases when evaluating diagnostic algorithms will help guide policy recommendations about the use of CXR and CAD as triage for molecular assays.

The findings of our systematic review suggest that additional research is needed before existing CAD technologies can be recommended for use. By highlighting important areas of uncertainty in the nascent evidence base, our review should help guide the direction of future evaluative studies of CAD4TB and other CAD programmes being developed (Advenio Tecnosys. Chandigarh, India; www.cad4globalhealth.com).[12,] Such studies should seek to minimise risks of bias through prospective designs and consecutive enrollment, pre-specified threshold scores, and microbiologically defined case definitions. Generalisability of the evidence base will be strengthened by evaluations undertaken in diverse settings and populations. Finally, there is also a need to discuss the potential limitations that could arise in using TB-centric CAD systems that cannot provide useful interpretations of other lung diseases.

### *Acknowledgments*

The authors would like to thank their medical librarian, N Torabi, for help with conducting one of the medical searches and establishing broad search terms.

FAK is a recipient of salary support from Fonds de Recherche du Québec Santé, Montreal, QC, Canada. MP holds a Tier 1 Canada Research Chair award.

Conflicts of interest: none declared.

### *References*

1 Pinto L M, Pai M, Dheda K, Schwartzman K, Menzies D, Steingart K R. Scoring systems using chest radiographic features for the diagnosis of pulmonary tuberculosis in adults: a systematic review. Eur Respir J 2012; 42: 480–494.

2 Samandari T, Bishai D, Luteijn M, et al. Costs and consequences of additional chest x-ray in a tuberculosis prevention program in Botswana. Am J Respir Crit Care Med 2011; 183: 1103–1111.

3 Chunhaswasdikul B, Kamolratanakul P, Jittinandana A, Tangcharoensathien V, Kuptawintu S, Pantumabamrung P. Anti-tuberculosis programs in Thailand: a cost analysis. Southeast Asian J Trop Med Public Health 1992; 23: 195–199.

4 Koppaka R, Bock N. How reliable is chest radiography? In: Toman's tuberculosis: case detection, treatment, and monitoring – questions and answers. 2nd ed. Geneva, Switzerland: World Health Organization, 2004: pp 51–60.

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

6 Hogeweg L, Sanchez C, Maduskar P, et al. Automatic detection of tuberculosis in chest radiographs using a combination of

textural, focal, and shape abnormality analysis. IEEE Trans Med Imaging 2015; 34: 2429–2442.
7 Philipsen R, Sánchez C, 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.
8 Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, van Ginneken B. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. Int J Tuberc Lung Dis 2013; 17: 1613–1620.
9 Muyoyeta M, Maduskar P, Moyo M, et al. The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia. PLOS ONE 2014; 9: e93757.
10 Breuninger M, van Ginneken B, Philipsen R H, et al. Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-saharan Africa. PLOS ONE 2014; 9: e106381.
11 Nathanson C M, Cuevas L E, Cunningham J, et al. The TDR tuberculosis specimen bank: a resource for diagnostic test developers. Int J Tuberc Lung Dis 2010; 14: 1461–1467.
12 Foundation for Innovative New Diagnostics. Digital radiology solutions for TB diagnosis in low- and middle-income countries. Technology Landscape Report. Geneva, Switzerland: FIND, 2015.

## APPENDIX

| Study authors | Risk of bias | | | | Applicability concerns | | |
|---|---|---|---|---|---|---|---|
| | Patient selection | Index test | Reference standard | Flow and timing | Patient selection | Index test | Reference standard |
| Maduskar[8] | U | L | L | U | H | H | L |
| Breuninger[10] | H | L | H | H | H | L | H |
| Muyoyeta[9] | L | L | L | L | U | U | L |
| Hogeweg[6] | H | L | U | U | U | U | L |
| Philipsen[7] | L | L | L | L | L | L | L |

L = low; H = high; U = unclear.

**Figure A** Quality assessment (QUADAS 2): risk of bias and applicability concerns; L = low; H = high; U = unclear.

## SEARCH STRATEGIES

*PUBMED*

((((‘Diagnosis, Computer-Assisted’[Mesh] OR ‘computer assisted diagnosis’ OR ‘computer aided diagnosis’ OR ‘radiographic image interpretation’ OR ‘computer’ OR ‘computer automated’)) AND (‘Tuberculosis, Pulmonary’[Mesh] OR TB OR ‘tuberculosis’ OR ‘pulmonary tuberculosis’ OR ‘pulmonary tuberculoses’ OR mycobacterium tuberculosis))) AND (sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic *[MeSH:noexp] OR diagnosis, differential[MeSH:noexp] OR diagnosis[Subheading:noexp])

*EMBASE/OVID*

1 tuberculosis/ or lung tuberculosis/
2 Pulmonary tuberculosis.mp.
3 (tuberculosis or TB or tuberculoses).ti,ab.
4 mycobacterium tuberculosis.mp. or Mycobacterium tuberculosis/
5 1 or 2 or 3 or 4
6 computer assisted diagnosis/ or computer assisted radiography/
7 (computer adj2 diagnosis).ti,ab.
8 (computer adj2 detection).ti,ab.
9 (computer adj2 screening).ti,ab.
10 (computer).ti,ab
11 radiographic image interpretation.mp.
12 automatic screening.mp.
13 6 or 7 or 8 or 9 or 10 or 11 or 12
14 sensitiv:.tw.
15 diagnostic accuracy.sh.
16 diagnostic.tw.
17 14 or 15 or 16
18 5 and 13 and 17
19 di.fs.
20 predict:.tw.
21 specificity.tw.
22 19 or 20 or 21
23 5 and 13 and 22

*SCOPUS*

( ( TITLE-ABS-KEY ( ‘tuberculosis’ ) ) OR ( TITLE-ABS-KEY ( ‘pulmonary tuberculosis’ ) ) OR ( TITLE-ABS-KEY ( ‘mycobacterium tuberculosis’ ) ) ) AND ( ( TITLE-ABS-KEY ( ‘computer assisted diagnosis’ ) ) OR ( TITLE-ABS-KEY ( ‘computer aided diagnosis’ ) ) OR ( TITLE-ABS-KEY ( ‘computer aided detection’ ) ) OR ( TITLE-ABS-KEY ( ‘computer aided screening’ ) ) OR ( TITLE-ABS-KEY ( ‘computer assisted detection’ ) ) OR ( TITLE-ABS-KEY ( ‘computer assisted screening’ ) ) OR ( TITLE-ABS-KEY ( ‘automatic screening’ ) ) ) AND ( ( ALL ( ‘specificity’ ) ) OR ( ALL ( ‘sensitivity’ ) ) OR ( ALL ( ‘diagnosis’ ) ) )

*ENGINEERING VILLAGE*

(((((((($Computer $assisted $diagnosis) WN ALL) OR (($computer $aided $diagnosis) WN ALL)) OR (($computer $aided $detection) WN ALL)) OR (($computer $aided $screening) WN ALL)) OR (($computer $assisted $detection) WN ALL)) OR (($computer $assisted $screening) WN ALL)) OR (($automatic $screening) WN ALL)) AND (1884–2015 WN YR)) AND ( ((($tuberculosis) WN ALL) OR (($pulmonary $tuberculosis) WN ALL)) OR (($mycobacterium $tuberculosis) WN ALL)) AND (1884–2015 WN YR))

Quality assessment (QUADAS 2): risk of bias and applicability concerns: See Figure A.

## RESUME

**OBJECTIF :** Revoir systématiquement l'exactitude diagnostique de la détection assistée par ordinateur (CAD) de la tuberculose pulmonaire (TBP) sur les radiographies pulmonaires (CXR) numérisées.

**SCHÉMA :** Nous avons recherché dans quatre bases de données les articles publiés entre janvier 2010 et décembre 2015 qui ont comparé la CAD de la CXR pour suspicion de TB pulmonaire à une référence microbiologique standard (frottis, culture ou réaction de polymérase en chaîne). Nous avons recueilli et résumé les données relatives à la conception de l'étude, au logiciel de CAD et à la précision diagnostique (sensibilité, spécificité, aire sous la courbe [AUC]).

**RÉSULTATS :** Nous avons inclus 5 sur 455 articles identifiés par notre recherche sur les bases de données. La prévalence de la TBP allait de 18% à 60% et celle de l'infection par le virus de l'immunodéficience humaine (VIH), de 33% à 68%. Tous les articles ont évalué CAD4TB, qui est le seul logiciel disponible dans le commerce. L'AUC allait de 0,71 à 0,84. Les paramètres du logiciel qui augmentaient la sensibilité réduisaient considérablement la spécificité et vice versa. Le risque de biais a été faible dans les études prospectives ($n = 2$) et élevé dans les études rétrospectives ($n = 3$).

**CONCLUSION :** L'évaluation de l'apport de la CAD à l'exactitude diagnostique est limitée par le petit nombre d'études, dont la majorité souffre d'importantes limites méthodologiques, la disponibilité et l'évaluation d'un seul programme logiciel et la possibilité limitée de généraliser les résultats à des contextes dans lesquels la TBP et le VIH sont moins prévalents. Des recherches supplémentaires sont requises.

## RESUMEN

**OBJETIVO:** Analizar de manera sistemática la precisión diagnóstica de la detección asistida por computador (CAD) de la tuberculosis pulmonar (TBP) a partir de las imágenes numéricas de las radiografías de tórax (CXR).

**MÉTODOS:** Se llevó a cabo una búsqueda en las bases de datos de artículos publicados entre enero del 2010 y diciembre del 2015 que compararan el diagnóstico de la TBP mediante un CAD de las CXR con los métodos microbiológicos de referencia (baciloscopia, cultivo y reacción en cadena de la polimerasa). Se recogieron y resumieron los datos sobre los métodos de los estudios, el programa CAD y la precisión diagnóstica (la sensibilidad, la especificidad y el área bajo la curva [AUC] en el análisis de eficacia diagnóstica).

**RESULTADOS:** Se incluyeron en el análisis cinco de los 455 artículos que se encontraron en las bases de datos. La prevalencia de TBP osciló entre 18% y 60%, y la prevalencia de infección por el virus de la inmunodeficiencia humana (VIH) entre 33% y 68%. Todos los artículos evaluaron el programa CAD4TB, que es el único programa informático de asistencia al diagnóstico de la TB disponible en el comercio. El AUC osciló entre 0,71 y 0,84. Las configuraciones del programa que mejoraron la sensibilidad disminuyeron de manera considerable la especificidad y viceversa. El riesgo de sesgo fue bajo en los estudios prospectivos ($n = 2$) y alto en los estudios retrospectivos ($n = 3$).

**CONCLUSIÓN:** Los datos probatorios sobre la precisión de la CAD de la TB por computador se encuentran limitados por el escaso número de estudios que la evalúan, de los cuales la mayoría adolece de limitaciones metodológicas, por la disponibilidad y la evaluación de un solo programa informático y porque la generalización del método se limita a los entornos donde la TBP y la infección por el VIH son menos prevalentes. Se precisan nuevas investigaciones en este campo.