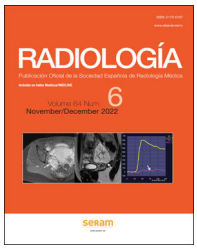
Radiología 66 (2024) 326–339

RADIOLOGÍA

www.elsevier.es/rx



ORIGINAL ARTICLES

Methodological evaluation of systematic reviews based on the use of artificial intelligence systems in chest radiography

J. Vidal-Mondéjar[a,*], L. Tejedor-Romero[a], F. Catalá-López[b,c,d]

[a] *Servicio de Medicina Preventiva, Hospital Universitario de La Princesa, Madrid, Spain*
[b] *Departamento de Planificación y Economía de la Salud, Escuela Nacional de Sanidad, Instituto de Salud Carlos III, Madrid, Spain*
[c] *Departamento de Medicina, Universidad de Valencia/Instituto de Investigación Sanitaria INCLIVA y CIBERSAM, Valencia, Spain*
[d] *Knowledge Synthesis Group, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

Received 6 December 2022; accepted 18 January 2023
Available online 26 July 2024

KEYWORDS
Artificial intelligence;
Methodology;
Chest X-ray;
Systematic review

Abstract
Introduction: In recent years, systems that use artificial intelligence (AI) in medical imaging have been developed, such as the interpretation of chest X-ray to rule out pathology. This has produced an increase in systematic reviews (SR) published on this topic. This article aims to evaluate the methodological quality of SRs that use AI for the diagnosis of thoracic pathology by simple chest X-ray.
Material and methods: SRs evaluating the use of AI systems for the automatic reading of chest X-ray were selected. Searches were conducted (from inception to May 2022): PubMed, EMBASE, and the Cochrane Database of Systematic Reviews. Two investigators selected the reviews. From each SR, general, methodological and transparency characteristics were extracted. The PRISMA statement for diagnostic tests (PRISMA-DTA) and AMSTAR-2 were used. A narrative synthesis of the evidence was performed. Protocol registry: Open Science Framework: https://osf.io/4b6u2/.
Results: After applying the inclusion and exclusion criteria, 7 SRs were selected (mean of 36 included studies per review). All the included SRs evaluated ''deep learning'' systems in which chest X-ray was used for the diagnosis of infectious diseases. Only 2 (29%) SRs indicated the existence of a review protocol. None of the SRs specified the design of the included studies or provided a list of excluded studies with their justification. Six (86%) SRs mentioned the use of PRISMA or one of its extensions. The risk of bias assessment was performed in 4 (57%) SRs. One (14%) SR included studies with some validation of AI techniques. Five (71%) SRs presented results in favour of the diagnostic capacity of the intervention. All SRs were rated c̈ritically lowf̈ollowing AMSTAR-2 criteria.

* Corresponding author.
E-mail address: jaime.vidal@salud.madrid.org (J. Vidal-Mondéjar).

Conclusions: The methodological quality of SRs that use AI systems in chest radiography can be improved. The lack of compliance in some items of the tools used means that the SRs published in this field must be interpreted with caution.

PALABRAS CLAVE
Inteligencia artificial;
Metodología;
Radiografía de tórax;
Revisión sistemática

Evaluación metodológica de las revisiones sistemáticas basadas en la utilización de sistemas de inteligencia artificial en radiografía de tórax

Resumen
Introducción: En los últimos años se han desarrollado sistemas que utilizan inteligencia artificial (IA) para estudiar distintos aspectos de la imagen médica, como la interpretación de la radiografía de tórax para descartar patología. Esto ha producido un aumento de las revisiones sistemáticas (RS) publicadas sobre este tema. Este artículo tiene como objetivo evaluar la calidad metodológica de las RS que utilizan IA para el diagnóstico de patología torácica mediante radiografía de tórax.
Material y métodos: Se seleccionaron RS que evaluaran el uso de sistemas de IA para la lectura automática de radiografía de tórax. Se realizaron búsquedas (desde el inicio hasta mayo de 2022) en: PubMed, EMBASE y Cochrane Database of Systematic Reviews. Dos investigadores seleccionaron los estudios. De cada RS, se extrajeron elementos generales, metodológicos y de transparencia de la presentación. Se utilizaron las guías PRISMA para pruebas diagnósticas (PRISMA-DTA) y AMSTAR-2. Se realizó una síntesis narrativa de la evidencia. Registro del protocolo: Open Science Framework: https://osf.io/4b6u2/.
Resultados: Tras aplicar los criterios de inclusión y exclusión, se seleccionaron 7 RS (media de 36 estudios incluidos por revisión). Todas las RS incluidas evaluaron sistemas de ''aprendizaje profundo'' en los que se utilizaba la radiografía de tórax para el diagnóstico de enfermedades infecciosas. Solo 2 (29%) RS indicaron la existencia de un protocolo. Ninguna RS especificó el diseño de los estudios incluidos ni facilitó una lista de estudios excluidos con su justificación. Seis (86%) RS mencionaron la utilización de PRISMA o alguna de sus extensiones. La evaluación del riesgo de sesgos se realizó en 4 (57%) RS. Una (14%) RS incluyó estudios con alguna validación de las técnicas de IA. Cinco (71%) RS presentaron resultados a favour de la capacidad diagnóstica de la intervención. Todas las RS obtuvieron la calificación ''críticamente baja'' siguiendo criterios AMSTAR-2.
Conclusiones: La calidad metodológica de las RS que utilizan sistemas de IA en radiografía de tórax es mejorable. La falta de cumplimiento en algunos ítems de las herramientas utilizadas hace que las RS publicadas en este campo deban interpretarse con cautela.

Introduction

Chest radiography is a widely used technique in clinical practice. It is easier to perform, less expensive and requires lower radiation doses than other more complex techniques, such as computed tomography (CT).[1,2] It has been estimated that between 1997–2007 alone, 3.6 billion diagnostic tests were performed, and of these 40% corresponded to chest radiographs.[3] Chest radiographs are interpreted by a wide range of medical professionals, and training is essential for readings to be accurate.[2] One of the current healthcare system challenges is the increasing demand for diagnostic tests, accompanied by a shortage of specialists that can read them, so it is not uncommon for diagnostic radiology services to struggle to read and report on all the tests performed.[4,5]

Among all the developments that could bring about major changes in diagnostic imaging, the most striking is artificial intelligence (AI).[6] AI is a field of computer science designed to mimic human intelligence with computer systems through an iterative comparison of complex patterns, usually at a speed and scale that exceeds human capabilities.[7] AI could have multiple applications in the field of medical imaging, including computer-aided diagnosis, detection of poor quality studies or the automatic selection of technical parameters by imaging systems.[8–11] AI systems need a significant volume of images for their training if they are to achieve an adequate level of accuracy. In recent years, several studies have sought to determine the number of samples needed to ensure that each algorithm functions correctly. This varies depending on the imaging study type.[11] Pub-

lished research on this topic has increased; however, many of these individual studies are low-quality in terms of their methodology and data reporting.[12]

Systematic reviews (SR) aim to bring together all empirical evidence in order to answer a specific research question. They use systematic and explicit methods, which are intended to minimise bias, thus providing more reliable results from which conclusions can be drawn.[13,14] When SRs are carried out properly, following methodological guidelines or standards, they provide a high level of evidence. A number of SRs have been published that relate to advances in diagnostic techniques that incorporate AI.[12,15–18] However, no published studies have assessed the methodological quality of published SRs of studies on AI-assisted chest radiography diagnosis. The main objective of this study is to assess the methodological quality of SRs of studies that use AI-based techniques for diagnosis by plain chest radiography.

Materials and methods

This SR examines the methodological quality of diagnostic test reviews.

Study design and protocol registration

A protocol was created and registered (*Open Science Framework*: https://osf.io/4b6u2/)) before commencement. During this study, there were no significant deviations from the predetermined objectives. This SR of methodological quality followed the recommendations of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 statement[19] (Appendix B p. 3–5 in Supplementary material).

Eligibility criteria

Articles were selected in line with criteria related to the study design/type, population group, tests carried out, results, type of publication and language.

Study types: SRs both with and without meta-analyses were included irrespective of study type (e.g. clinical trials and observational studies). We included SRs that explicitly stated the methods used for identifying studies (e.g. a search strategy) and for selecting studies (e.g. eligibility criteria) and those that described their synthesis methods (with or without quantitative data).

Population groups: SRs should include studies performed on healthy individuals and/or individuals with any sign of chest disease undergoing chest radiography, without exclusion by ethnicity, sex or age.

Intervention and comparator: SRs should evaluate the use of an AI system, whether a computer-aided diagnosis system or a deep learning or machine learning system, for the diagnostic interpretation of plain chest radiographs. Both conventional and digital radiographs were included. SRs whose main purpose was chest radiograph assessment were included. SRs of multiple diagnostic modalities (for example, CT, ultrasound…) which evaluated AI systems were excluded even if they included plain chest radiography. The comparator or reference test should at least mention the usual or standard clinical management practice (e.g. interpretation by a radiologist, or microbiological confirmation in the case of infectious diseases).

Findings of interest: The SRs should relate to any chest radiograph interpretations and the studies included in them should evaluate the diagnostic precision or reliability of the tests (for example, sensitivity, specificity or predictive values).

Length of follow-up: no threshold was set for the length of follow-up.

Publication status: both published SRs and SRs awaiting publication (pre-publication) were evaluated.

Languages: SRs written in English or Spanish were included.

Information sources and search strategy

To find the articles, we conducted an exhaustive search of the main databases (from their start date to 26 May 2022): MEDLINE (through PubMed), EMBASE, and the Cochrane Database of Systematic Reviews. An experienced documentalist helped design the search strategies (CA-FM, Biblioteca Nacional de Ciencias de la Salud, Instituto de Salud Carlos III) which included keywords related to AI, chest radiography and SR/meta-analysis (Appendix B p. 6 and 7 of the Supplementary material). We also checked Google Scholar and the bibliographic references of potentially eligible articles, contacting authors if additional information was needed, in order to increase the sensitivity of the searches.

Study selection

Two researchers (J. V-M and L. T-R) carried out the SR selection following the inclusion and exclusion criteria. Any discrepancies were discussed and resolved with a third researcher (F. C-L). The Rayyan® software (Rayyan Systems Inc., Cambridge, USA)[20] was used for this purpose, and duplicate articles were deleted.

Data extraction

Two researchers (J. V-M and L. T-R) independently extracted relevant data from the included SRs, as follows:

General characteristics of the SRs: first author and year of publication; country of corresponding author; journal name and impact factor (according to Journal Citation Reports; 2021); number of databases used and names (for example, PubMed, EMBASE, Scopus and others); mention of the existence of a protocol (yes/no), and if yes, where it can be accessed (for example, PROSPERO); description of the tools used, both for reporting (for example, PRISMA statement[19]) and for methodological quality assessment (for example, A MeaSurement Tool to Assess systematic Reviews 2 [AMSTAR-2][21,22]); adequate/inadequate use of PRISMA (according to criteria defined by Caulley et al.[23]) and the reporting of a completed checklist; mention of an instrument to assess the risk of bias in studies (for example, QUality Assessment of Diagnostic Accuracy Studies [QUADAS-2]); and source of funding.

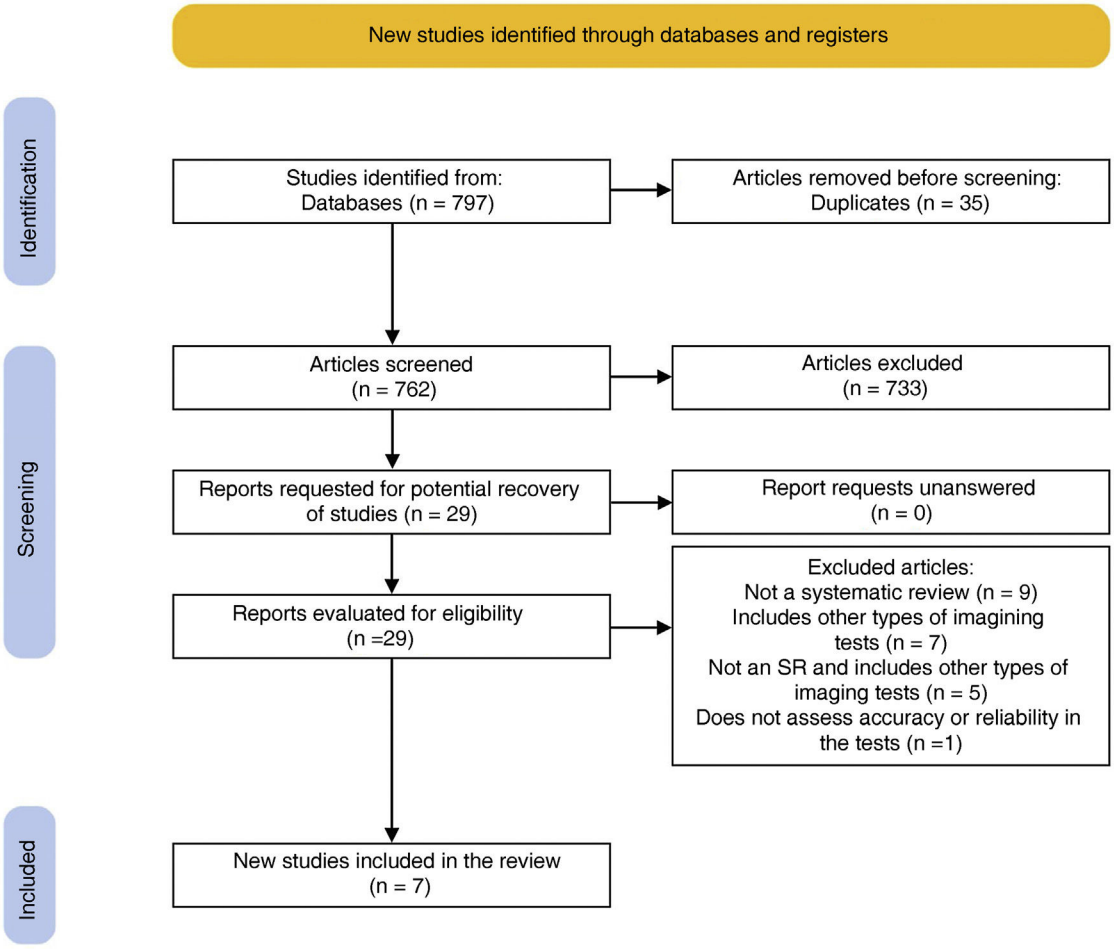


Figure 1 Study selection flowchart.

Specific characteristics of the SRs: description of the intervention (e.g. name of the algorithm evaluated, type of architecture, sources of training data and number of images), description of the comparator (e.g. diagnostic confirmation as standard practice and microbiological tests), number and design of included studies (e.g. randomised trials and observational studies), description of the characteristics of the participants in the included studies, whether validation was performed and what type, measure of accuracy used (e.g. sensitivity and specificity).

Synthesis methods used in SRs: type of synthesis (e.g. narrative/qualitative and quantitative/meta-analysis), analysis model used if applicable, reporting of pooled data and its 95% confidence interval, and additional analyses (e.g. subgroup analysis and meta-regression).

Qualitative results and conclusions reported in the SRs: favourable if the evaluated AI system was clearly the recommended option (e.g. mentioned as 'effective', 'beneficial', 'improves diagnostic accuracy', 'promising technique'), unfavourable if the conclusions were clearly negative (e.g. 'not effective', 'unlikely to be beneficial', 'does not improve diagnosis') and neutral or inconclusive when the tool of interest was not superior to the comparator or when the conclusions were expressed with a high degree of uncertainty.

Evaluation of reporting transparency and methodological quality

Two researchers (J. V-M and L. T-R) assessed the transparency of reporting and methodological quality of the included SRs. AMSTAR-2 was used to assess methodological quality.[21,22] AMSTAR-2 presents a checklist with short answer options ('yes', 'partial yes' and 'no') that evaluates different relevant domains of the SRs.[21,22] These domains are divided into seven 'critical' and nine 'non-critical' domains according to the level that they are expected to affect the validity of the results. The results of each item can be 'yes' for complete adherence, 'partial yes' when an item adheres under certain variable circumstances or partially adheres and 'no' for failure to meet the item criteria. There are four rating or confidence options: 'high' (no critical weaknesses and maximum of one non-critical weakness), 'moderate' (no critical weaknesses and two or more non-critical weaknesses), 'low' (maximum of one critical weakness regardless of the number of non-critical weaknesses) and 'critically low' (two or more critical weaknesses).[21,22]

The tool used to assess transparency in the reporting of the included SRs was the PRISMA Diagnostic Test Accuracy (DTA) statement.[24] The PRISMA-DTA[24] statement was published in January 2018, and includes a 27-item checklist

designed to improve the completeness of reporting on the methods and results of SRs of studies on diagnostic tests.

Data analysis

We produced a narrative summary of the main characteristics of the included SRs, and all extracted data were presented in evidence tables. General, methodological and reporting transparency elements were presented for each SR. A descriptive analysis included frequencies and percentages to express the results obtained from the AMSTAR-2[21,22] and PRISMA-DTA[24] checklists.

Results

Search results and selection of included systematic reviews

From the searches, 797 reviews were identified, and 35 duplicates eliminated. Of these, 733 were excluded as they were considered irrelevant after reading the title and abstract. In the end, 29 articles were evaluated by full-text reading. After excluding 22 articles (Appendix B see p 8 and 9 of the Supplementary material), seven SRs[16–18,25–28] were included (Fig. 1).

General characteristics of the systematic reviews

The general characteristics of the included SRs are set out in Table 1. The seven SRs mention the databases used for the searches, and all seven consulted at least three (range: three to five databases). Only two (29%) SRs reported a descriptive and accessible protocol. Six (86%) mention that they use PRISMA or one of its extensions, but only one included a completed checklist (in its annex). Risk of bias assessment was carried out in four (57%) SRs, mainly using the QUADAS-2 tool (n = 3; 43%).

Specific characteristics of the systematic reviews

Table 2 describes the specific characteristics of the SRs and studies included. The seven SRs evaluated studies on AI systems using deep learning techniques. Two (29%) SRs also mention other machine-learning tools. Seven assessed the capability of AI systems to diagnose infectious thoracic diseases (for example, tuberculosis: n = 4; 57%; pneumonia: n = 2; 29%; and COVID-19: n = 1). One SR[17] was based on the automatic reading of chest radiographs in children. Only two (29%) SRs[26,27] clearly defined how many images were used (for example, both for the training and for the evaluation of AI systems). Five (71%) described the information sources used to obtain the radiological images. Four (57%) SRs described the comparator used (for example, microbiological reference standard of the infectious disease and/or clinical criteria).

All seven reported the number of studies included (mean = 36 studies and range: 4–62 studies per review). No SRs described the study designs of the included studies. Three (43%) mentioned some general characteristics about the study participants. Seven described the measures of accuracy used in the tests, such as sensitivity, specificity or area under the curve (AUC) (Table 2). Only one SR[17] reported that in some of the included studies, validation of the AI systems had been carried out, whether internal or external. In the other SRs (n = 6; 86%) no references are made to any validation.

Table 3 sets out descriptions of the synthesis methods used. No meta-analysis was carried out in five of the SRs (71%). One SR with meta-analysis[27] compared the capability of the AI systems to diagnose pneumonia to its ability to distinguish between viral and bacterial pneumonia (e.g. sensitivity = 0.98; specificity = 0.94 and AUC = 0.99). Another SR with meta-analysis and meta-regression[18] evaluated the capability of different AI systems to assess pulmonary tuberculosis (e.g. AUC = 0.83–0.85 for different AI software systems; specificity = 0.54–0.60).

Results and qualitative conclusions of the systematic reviews

Five (71%) SRs presented 'favourable' results for the use of AI to make the diagnosis of infectious diseases with chest radiograph more accurate and reliable. Only two SRs[16,18] were somewhat more critical when it came to reporting and interpreting these results (Table 4).

Results of the evaluation of reporting transparency and methodological quality

With regard to evaluating methodological quality, seven SRs were deemed 'critically low' according to the AMSTAR-2 criteria (Appendix B see p. 10 and 11 of the Supplementary material). For example, only two SRs[18,26] obtained 'partial yes' responses (in reference to the presence of a protocol, previous description of the steps to be taken in the review and justification of any significant deviation from the protocol). No SRs adhered to critical domain 7 (presentation of a list of studies excluded with the reasons for each exclusion).

The level of transparency and quality in the reporting of the included SRs varied according to the criteria of the PRISMA statement.[19,24] For example, four SRs (57%) did not adequately describe the search strategy (item 8), three (43%) did not describe the risk of bias assessment (item 13) and four (57%) did not adequately report on how the results synthesis was carried out (item 14). Only one SR[16] adequately described the study characteristics. Only two adequately reported the risk of bias (item 19). In Appendix B of the Supplementary material (p12 and 13) the PRISMA-DTA[24] statement checklist was filled out in detail for each SR. When we examined whether or not each of the SR that mention PRISMA used it appropriately, we found that only two[17,27] appear to have used it properly. One SR used it inadequately[16] and there were doubts around its use in the others (three SRs[25,26,28]).

Discussion

The main finding of this research is that methodology varies in quality among published SRs of studies which evaluate

Table 1 General characteristics of the included systematic reviews.

Author and year of publication	Country of the corresponding author	Name of journal	Impact factor (2020)	Databases used, number (name)	Existence of protocol and access to it	Methodological reporting tools	Completed PRISMA checklist	Risk of bias assessment instrument	Funding source
Ghaderzadeh et al.,[25] 2021	Iran	Biomed Res Int	3.411	4 (Scopus, Elsevier ScienceDirect, PubMed and Web of Science)	No	PRISMA 2009	No	Modified CHARM checklist	Public
Harris et al.,[26] 2019	Canada	PLoS One	3.240	4 (PubMed, MEDLINE, EMBASE and Scopus)	Yes, PROSPERO	PRISMA 2009	Yes	QUADAS-2	None
Li et al.,[27] 2020	China	Comput Biol Med	4.589	5 (PubMed, Embase, Scopus, Web of Science and Google Scholar)	No	PRISMA 2009	No	No	None
Oloko-Oba et al.,[28] 2022	South Africa	Front Med	5.093	4 (Scopus, IEEE Xplore, Web of Science and PubMed)	No	PRISMA 2009	No	No	Not described
Padash et al.,[17] 2022	Canada	Pediatr Radiol	2.505	3 (PubMed, EMBASE and Web of Science)	No	PRISMA-DTA	No	No	Not described
Pande et al.,[16] 2016	Canada	Int J Tuberc Lung Dis	2.373	4 (PubMed, EMBASE, Scopus and Engineering Village)	No	No	No	QUADAS-2	Public
Tavaziva et al.,[18] 2021	Canada	Clin Infect Dis	9.079	5 (MEDLINE, EMBASE, PubMed, Scopus and Engineering Village)	Yes, PROSPERO	PRISMA-IPD	No	QUADAS-2	Public

CHARMS: CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies; IEEE: Institute of Electrical and Electronics Engineers; PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses; PRISMA-DTA: Preferred Reporting Items for Systematic reviews and Meta-Analyses Diagnostic Test Accuracy; PRISMA-IPD: Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data; PROSPERO: International Prospective Register of Ongoing Systematic Reviews; QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2.

Table 2 Specific characteristics of the included systematic reviews.

Author and year of publication	Description of the intervention	Description of the comparator	Number and design of studies included in the review	Description of the participant characteristics	Existence of validation and type (internal/external)	Measure of accuracy used
Ghaderzadeh et al.,[25] 2021	1. ML- and DL-based AI methods in chest radiographs in patients with COVID-19 2. Differentiate AI systems in terms of objective (classify, detect lesions, segmentation...) 3. No. of images in training: Not defined 4. No. of images tested: Not defined. 5. Description of databases used: Yes	No complete description of comparator	60 studies Does not specify design	General patient characteristics are collected: No Other data that describe participants: Differentiate between dichotomous studies (COVID yes/no) vs. studies with more variables No. of participants: 86—13975 Disease studied: COVID-19	No/No	Sensitivity, specificity, accuracy, AUC, F1-score, recall
Harris et al.,[26] 2019	1. ML-based (n = 46) and DL-based (n = 7) software in chest radiography in patients with tuberculosis. 2. Multiple AI-based algorithms (CAD4TB, many others not defined...) 3. No. of images in training: 18 - 60989. 4. No. of images tested: 30 - 37475. 5. Description of databases used: Yes	Microbiological reference standard or start of treatment: 17 Human reader: 39	53 studies Does not specify design	General patient characteristics are collected: Yes Other data that describe participants: Triage studies vs. screening studies No. of participants: 161—17006 Disease studied: tuberculosis	No/No	AUC, sensitivity, specificity, true positives, false positives, false positive ratios, true negatives, false negatives

Table 2 (*Continued*)

Author and year of publication	Description of the intervention	Description of the comparator	Number and design of studies included in the review	Description of the participant characteristics	Existence of validation and type (internal/external)	Measure of accuracy used
Li et al.,[27] 2020	1. DL-based AI methods 2. Disease diagnosis vs. normality and differentiation between pneumonia types (viral and bacterial). 3. No. of images in training: 2000 – 25648. 4. No. of images tested: 300 – 1928 5. Description of databases used: Yes	No complete description of comparator	15 studies Does not specify design	General patient characteristics are collected: No Other data that describe participants: No No. of participants: Not defined Disease studied: pneumonia, viral and bacterial	No/No	AUC, sensitivity, specificity, true positives, false positives, negative likelihood ratio, positive likelihood ratio, true negatives, false negatives, diagnostic odds ratio
Oloko-Oba et al.,[28] 2022	1. DL-based AI methods 2. DL techniques for diagnosing tuberculosis 3. No. of images in training: Not defined 4. No. of images tested: Not defined 5. Description of databases used: Yes	No complete description of comparator, one of the included studies mentions that the comparators are the doctors	62 studies Does not specify design	General patient characteristics are collected: No Other data that describe participants: Not defined No. of participants: Not defined Disease studied: tuberculosis	No/No	AUC, sensitivity, specificity, accuracy, recall, F1-score

Table 2 (*Continued*)

Author and year of publication	Description of the intervention	Description of the comparator	Number and design of studies included in the review	Description of the participant characteristics	Existence of validation and type (internal/external)	Measure of accuracy used
Padash et al.,[17] 2022	1. Description of databases of paediatric radiography 2. Use of DL for diagnosis of pneumonia and other diseases 3. No. of images in training: Not defined 4. No. of images tested: Not defined 5. Description of databases used: Yes	Radiological or clinical assessment	55 studies Does not specify design	General patient characteristics are collected: Yes Other data that describe participants: Age groups No. of participants: No Disease studied: Pneumonia (most studies), CF, RDS, bronchitis/bronchiolitis, pneumothorax	There is internal and external validation in some of the studies	AUC, sensitivity, specificity, accuracy
Pande et al.,[16] 2016	1. DL-based AI methods in radiography. 2. DL algorithm (CAD4TB for diagnosing tuberculosis 3. No. of images in training: Not defined 4. No. of images tested: Not defined 5. Description of databases used: No	Microbiological reference standard Culture + clinical criteria (1)	5 studies Does not specify design	General patient characteristics are collected: Partial Other data that describe participants: Country, HIV % No. of participants: 161—894 Disease studied: Tuberculosis	No/No	AUC, Sensitivity, Specificity
Tavaziva et al.,[18] 2021	1. DL-based AI methods in radiography 2. Use of various DL algorithms for diagnosing tuberculosis 3. No. of images in training: Not defined 4. No. of images tested: Not defined 5. Description of databases used: No	Microbiological reference standard (Sputum, culture or PCR) In a later analysis, comparison with human reader	4 studies Does not specify design	General patient characteristics are collected: Yes Other data that describe participants: HIV status, culture result... No. of participants: 342—2298 Disease studied: Tuberculosis	No/No	AUC, sensitivity, specificity, positive likelihood ratio, negative likelihood ratio

AI: artificial intelligence; AUC: area under the curve; CF: cystic fibrosis; DL: deep learning; DRS: Respiratory distress syndrome; ML: machine learning; PCR: polymerase chain reaction; HIV: human immunodeficiency virus.

Table 3 Synthesis methods applied in the included systematic reviews.

First author and year of publication	Type of synthesis	Model of analysis used	Reporting on pooled data of the meta-analysis	Additional analyses
Ghaderzadeh et al.,[25] 2021	Qualitative (narrative) and quantitative (without meta-analysis)	N/A	N/A	Sensitivity and specificity are compared with other studies
Harris et al.,[26] 2019	Qualitative (narrative) and quantitative (without meta-analysis)	N/A	N/A	AUC according to the study (development/clinical) and by type (DL/ML)
Li et al.,[27] 2020	Qualitative (narrative) and quantitative (with meta-analysis)	Not described	Sensitivity: 0.98 (95% CI: 0.96–0.99), Specificity: 0.94 (95% CI: 0.90–0.96) DOR: 718.13 (95% CI: 288.45–1787.93) Area under the ROC curve: 0.99 (95% CI: 0.98–100) PLR: 96.1 (95% CI: 96.1–97.7) NLR: 0.02 (95% CI: 0.01–0.04)	Studies with data on viral and bacterial pneumonia are analysed; Sensitivity: 0.89 (95% CI: 0.79–0.94) Specificity: 0.89 (95% CI: 0.78–0.95) DOR: 66.14 (95% CI: 17.34–252.37) Area under the ROC curve: 0.95 (0.93–0.97) PLR: 8.34 (95% CI: 3.75–18.55) NLR: 0.13 (95% CI: 0.06–0.26)
Oloko-Oba et al.,[28] 2022	Only qualitative (narrative)	N/A	N/A	Not performed
Padash et al.,[17] 2022	Only qualitative (narrative)	N/A	N/A	Not performed
Pande et al.,[16] 2016	Only qualitative (narrative)	N/A	N/A	Not performed
Tavaziva et al.,[18] 2021	Qualitative (narrative) and quantitative (with meta-analysis of individual patient data)	Random-effects model	AUC CAD4TBv6, 0.83 (95% CI: 0.82–0.84); Lunit, 0.83 (95% CI: 0.79–0.86); qxRv2, 0.85 (95% CI: 0.83–0.88) Specificity with 90% sensitivity: CAD4TBv6, 0.57 (95% CI 0.52–0.62); Lunit, 0.54 (95% CI: 0.45–0.63; qXRv2, 0.60 (95% CI: 0.52–0.69)	Analysis of subgroups (sex, HIV, sputum, history of tuberculosis and age); Meta-regression; post-hoc analysis compared to human readers

AUC: area under the curve; CI: confidence interval; DL: deep learning; DOR: diagnostic odds ratio; HIV: human immunodeficiency virus; ML: machine learning; NLR: negative likelihood ratio; PLR: positive likelihood ratio; ROC: receiver operating characteristic.

the use of AI in chest radiographs. Moreover, there is a lack of adherence to many of the items proposed in the methodological guidelines and standards used. Methodological tools such as AMSTAR-2[21,22] enable the identification of elements that affect the quality of the way the review is conducted, a key aspect when it comes to interpreting and evaluating the potential applicability of the results.

An SR can be considered to have 'high' methodological quality (according to the AMSTAR-2 criteria[21,22]) when it obtains a favourable result in all aspects evaluated. However, various critical domains are not adhered to in the SRs we analysed. For example, one of the most important aspects of the design of an SR is the elaboration and registration of a protocol which establishes—a priori—which methods will be used. AMSTAR-2[21,22] refers to these aspects and so it is striking that only two SRs registered a response of 'partial yes' as this can significantly compromise the methodological quality of these studies. Another key aspect

Table 4 Reporting of qualitative results in included systematic reviews.

First author and year of publication	Result/conclusion of the SR authors on the use of AI	Qualitative result
Ghaderzadeh et al.,[25] 2021	X-rays equipped with AI can serve as a tool to screen the cases requiring CT scans. The use of this tool does not waste time or impose extra costs, has minimal complications, and can thus decrease or remove unnecessary CT slices and other healthcare resources.	In favour of the intervention
Harris et al.,[26] 2019	We conclude that CAD programs are promising, but the majority of work thus far has been on development rather than clinical evaluation. We provide concrete suggestions on what study design elements should be improved.	In favour of the intervention
Li et al.,[27] 2020	DL indicated high accuracy performance in classifying pneumonia from normal CXR radiographs and also in distinguishing bacterial from viral pneumonia. However, major methodological concerns should be addressed in future studies for translating to the clinic.	In favour of the intervention
Oloko-Oba et al.,[28] 2022	We conclude that CAD systems are promising in tackling the challenges of the TB epidemic and made recommendations for improvement in future studies.	In favour of the intervention
Padash et al.,[17] 2022	Classification of chest radiographs as pneumonia was the most common application of AI, evaluated in 65% of the studies. Although many studies report high diagnostic accuracy, most algorithms were not validated on external datasets. Most AI studies for paediatric chest radiograph interpretation have focused on a limited number of diseases, and progress is hindered by a lack of large-scale paediatric chest radiograph datasets.	Neutral
Pande et al.,[16] 2016	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.	Neutral
Tavaziva et al.,[18] 2021	For CAD CXR analysis to be implemented as a high-sensitivity tuberculosis rule-out test, users will need threshold scores identified from their own patient populations and stratified by HIV and smear status.	In favour of the intervention

AI: artificial intelligence; CT: computed tomography.

of conducting an SR is the inclusion of a list of excluded studies (with the justification or reasons for exclusion). The absence of this information in the analysed SRs could compromise the quality of an SR as we cannot rule out possible selection biases in the studies. There are other aspects whose interpretation could be more complex when it comes to assessing methodological quality. For example, AMSTAR-2[21,22] specifies that the types of studies included in the review should be made clear; however, only one of the analysed SRs provided a detailed and clear description of the type of studies included. One possible reason to explain the absence of these definitions is that, at least in some cases, the researchers felt that it was already evident that they were examining studies that assessed diagnostic capabilities. Future reviews should give detailed and comprehensive descriptions of the types of studies included (e.g. experimental or observational) prior to and during the review.

In contrast to AMSTAR-2,[21,22] publication guidelines such as PRISMA,[19] and its extension, PRISMA-DTA,[24] focus on helping achieve complete and transparent reporting of the different sections of an SR. And yet, a simple reference to the PRISMA statement[19] does not necessarily mean that the authors have followed the guidelines correctly or that the SR has been carried out in a rigorous and transparent manner. When we examined the authors' use of the PRISMA statement in the included SRs, of the six SRs that cited it, only two appear to have used it in an appropriate manner. These results concur with those of other recent publications[23] which have identified that a significant number of research articles do not use these kinds of guidelines properly. It therefore seems necessary to promote training on these principal publication guidelines (such as the PRISMA statement[19,24]), both for SR authors and for reviewers and journal editors who publish them.

All sections of the methodological guides and standards must be adhered to in order to facilitate the reading and understanding of research. It is also essential to enable a quick and critical assessment as to whether or not this information has been included. The wide variability in the way the methodology is reported in the included SRs is striking, as is the fact that in some cases, the measures of accuracy to be used in the analysis are not shared a priori, or that an explanation of how the results of the included studies are going to be synthesised is missing. It is also worth highlighting that the results sections often lack data on study characteristics, while reporting on risk of bias assessment and individual study results are limited. As these tools are intended to make the reporting of methods and results

clearer, the absence of many of their required sections clearly limits the reporting transparency of the published SRs.

Several challenges can arise when designing this kind of study, or when analysing the data or interpreting the results of research that evaluates AI in medical imaging.[16] Most of the analysed SRs seem to favour the use of AI systems for improving the capacity to diagnose infectious diseases by chest radiography. Some cases also report limitations in the results but most do not provide a careful interpretation of the results or a qualitative evaluation of the results that favour the use of AI systems. When the diagnostic accuracy parameters are reported without providing a detailed and comprehensive description of the comparator, it is difficult to interpret the results of the primary studies and this can lead to errors. For example, when comparing the diagnostic accuracy of an AI software with a human reader (not necessarily a radiologist), what is being compared is the interpretation capacity. However, when you compare it with disease parameters such as microbiological references, you are comparing the accuracy in diagnosing the disease. Likewise, in almost all the included SRs, no clear definition of the comparator was provided; therefore, it is difficult to interpret the results and extrapolate them to real-life clinical situations. There is reason to suggest that many of the chest radiographs in the seven SRs were not interpreted by radiologists who are pulmonary specialists. If this is the case, this introduces a potential source of error as it has been shown that involving radiologists improves the quality of image interpretation (both because of their experience and because they have different equipment at their disposal including diagnostic viewers and screens with better spatial resolution). This means that the success of an AI tool may have been overstated. Likewise, in the case of chest radiographs (rather than other modalities such as CT or MRI which are generally interpreted by radiologists) the fact that no clear description is given of the human reader (radiologist vs. other doctors) could have a huge impact on the interpretation (and extrapolation) of the results. The non-existence of a comparator description could be one of the reasons why other recent articles[29] report that 'non-axial' modalities (such as chest radiograph or ultrasound) present a significantly higher risk of bias in the 'reference standard' domain than 'axial' modalities (CT and MRI).

Reporting of data related to these aspects and other characteristics (such as radiography databases, procedures for obtaining images for training, validation and comparative purposes) could be improved in many of the studies included in the analysed SRs. Some standard bodies for diagnostic test studies are currently developing specific versions that integrate the use of AI. For example, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) have developed an AI version: STARD-AI.[30] Moreover, the recently published Checklist for Artificial Intelligence in Medical Imaging[31] could help improve reporting transparency and quality in this type of study. There are currently no specific tools that assess the risk of bias in SRs of AI studies. It would be interesting to develop and apply these tools in future research. Moreover, future research should also look at other AI-related ethical and legal aspects. For example, under current legislation, the General Data Protection Regulation (GDPR) gives EU patients the right to an explanation of all decisions made by an algorithm. Future studies showing the usefulness of AI systems will have to be able to explain how they have come to their conclusions (transparency) and they will also have to obtain prospective informed content from patients to be able to use and exploit their medical images. On the other hand, several studies have not examined the risk of bias and, when they have, they have provided very little information on this assessment, both with regard to the study itself and the included databases. A number of the included SRs have compared some of the review results with data obtained from other research in order to arrive at some of the research conclusions. This could cause a methodological problem due to the high risk of bias when using data from studies that have followed different methodologies to obtain results and which are not integral to the specific study.

This methodological SR has been carried out following a protocol created a priori, with no significant deviations. We have sought to be transparent when reporting on both the study selection and data extraction. Furthermore, we evaluated all the main aspects of methodological quality twice, and the researchers worked out any potential discrepancies following the main guidelines and standards on methodology (PRISMA[19,24] and AMSTAR-2[21,22]). However, it is worth mentioning a few limitations to this study. Firstly, this evaluation has only dealt with SRs of a specific diagnostic modality, namely chest radiograph. This means that it is possible that the findings are not applicable to other fields of AI-based medical imaging. On the other hand, it is important to highlight that the tool that provides the most information on the methodological quality of SRs (AMSTAR-2[21,22]) may seem highly demanding, as all SRs report critical weaknesses. It may be that the results would have been different had we applied other tools to assess methodological quality and the risk of bias of SRs (for example, the ROBIS scale[32]), although these differences would be minor.[33] Finally, this study evaluated literature published in biomedical journals indexed in the main databases. Therefore, other SRs may be missing which have not yet been published or which are written in other languages (in particular, Chinese).

Conclusions

The results of this study indicate that the methodological quality of the SRs of studies that use AI systems in chest radiographs could be improved, with many having significant weaknesses. Non-compliance with a considerable number of the elements or characteristics analysed means that the SRs published in this field should be interpreted with caution. Improving methodological quality and reporting transparency of these SRs could make it easier to interpret the results obtained and to transfer them successfully to patient care.

Author contributions

J. Vidal-Mondéjar designed the study with help from L. Tejedor-Romero and F. Catalá-López. J. Vidal-Mondéjar and L. Tejedor-Romero gathered the data. J. Vidal-Mondéjar carried out the analysis. J. Vidal-Mondéjar, L. Tejedor-Romero and F. Catalá-López interpreted the data. J. Vidal-Mondéjar wrote the first draft of the manuscript. L. Tejedor-Romero

and F. Catalá-López made comments and critical revisions to the different versions of the text and edited the final version. F. Catalá-López supervised the study. All authors have read and approved the final version.

Funding

The authors declare that they have not received any funding for this work. F. C-L received grants from the Instituto de Salud Carlos III/CIBERSAM.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.rxeng.2023.01.015.

References

1. Sá dos Reis C, Pires-Jorge JA, York H, Flaction L, Johansen S, Maehle S. Curricula, attributes and clinical experiences of radiography programs in four European educational institutions. Radiography. 2018;24:e61–8, http://dx.doi.org/10.1016/j.radi.2018.03.002.
2. Jokerst C, Chung JH, Ackman JB, Carter B, Colletti PM, Crabtree TD, et al. ACR Appropriateness Criteria® acute respiratory illness in immunocompetent patients. J Am Coll Radiol. 2018;15:S240–51, http://dx.doi.org/10.1016/j.jacr.2018.09.012.
3. World Health Organization. In: Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk; 2016 [Accessed 11 January 2023]. Available from: https://apps.who.int/iris/bitstream/handle/10665/205033/9789241510349_eng.pdf
4. Kim J, Kim KH. Measuring the effects of education in detecting lung cancer on chest radiographs: utilization of a new assessment tool. J Cancer Educ. 2019;34:1213–8, http://dx.doi.org/10.1007/s13187-018-1431-8.
5. Faculty of Clinical Radiology. In: Standards for the communication of radiological reports and fail-safe alert notification; 2016 [Accessed 11 January 2023]. Available from: https://www.rcr.ac.uk/system/files/publication/field_publication_files/bfcr164_failsafe.pdf
6. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.
7. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. JAMA. 2019;321:31–2, http://dx.doi.org/10.1001/jama.2018.18932.
8. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10, http://dx.doi.org/10.1038/s41568-018-0016-5.
9. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. In: Lee G, Fujita H, editors. Deep Learning in medical image analysis. Challenges and applications. Cham: Springer International Publishing; 2020. p. 3–21.
10. Syed A, Zoga A. Artificial intelligence in radiology: Current technology and future directions. Semin Musculoskelet Radiol. 2018;22:540–5, http://dx.doi.org/10.1055/s-0038-1673383.
11. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers—From the radiology editorial board. Radiology. 2020;294:487–9, http://dx.doi.org/10.1148/radiol.2019192515.
12. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health. 2019;1:e271–97, http://dx.doi.org/10.1016/S2589-7500(19)30123-2.
13. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Chichester: John Wiley & Sons; 2019.
14. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: A cross-sectional study. PLoS Med. 2016;13:e1002028, http://dx.doi.org/10.1371/journal.pmed.1002028.
15. Kriza C, Amenta V, Zenié A, Panidis D, Chassaigne H, Urbán P, et al. Artificial intelligence for imaging-based COVID-19 detection: Systematic review comparing added value of AI versus human readers. Eur J Radiol. 2021;145:110028, http://dx.doi.org/10.1016/j.ejrad.2021.110028.
16. 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;20:1226–30, http://dx.doi.org/10.5588/ijtld.15.0926.
17. Padash S, Mohebbian MR, Adams SJ, Henderson RDE, Babyn P. Pediatric chest radiograph interpretation: How far has artificial intelligence come? A systematic literature review. Pediatr Radiol. 2022;52:1568–80, http://dx.doi.org/10.1007/s00247-022-05368-w.
18. Tavaziva G, Harris M, Abidi SK, Geric C, Breuninger M, Dheda K, et al. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. Clin Infect Dis. 2022;74:1390–400, http://dx.doi.org/10.1093/cid/ciab639.
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021:71, http://dx.doi.org/10.1136/bmj.n71.
20. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5:210, http://dx.doi.org/10.1186/s13643-016-0384-4.
21. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017:j4008, http://dx.doi.org/10.1136/bmj.j4008.
22. Cinquini M, Moschetti I, Minozzi S. Assessing the methodological quality of systematic review: the AMSTAR II-DTA extension. In: Abstracts of the 26th Cochrane Colloquium, Santiago, Chile. Cochrane Database Syst Rev. 2020; 1 Suppl 1), http://dx.doi.org/10.1002/14651858.CD201901.
23. Caulley L, Catalá-López F, Whelan J, Khoury M, Ferraro J, et al. Reporting guidelines of health research studies are frequently used inappropriately. J Clin Epidemiol. 2020;122:87–94, http://dx.doi.org/10.1016/j.jclinepi.2020.03.006.
24. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA. 2018;319:388, http://dx.doi.org/10.1001/jama.2017.19163.

25. Ghaderzadeh M, Aria M, Asadi F. X-ray equipped with artificial intelligence: Changing the COVID-19 diagnostic paradigm during the pandemic. BioMed Res Int. 2021;2021:1–16, http://dx.doi.org/10.1155/2021/9942873.
26. Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, 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, http://dx.doi.org/10.1371/journal.pone.0221339.
27. Li Y, Zhang Z, Dai C, Dong Q, Badrigilan S. Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis. Comput Biol Med. 2020;123:103898, http://dx.doi.org/10.1016/j.compbiomed.2020.103898.
28. Oloko-Oba M, Viriri S. A systematic review of deep learning techniques for tuberculosis detection from chest radiograph. Front Med. 2022;9:830515, http://dx.doi.org/10.3389/fmed.2022.830515.
29. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, Darzi A. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: A meta-research study. NPJ Digit Med. 2022;5:11, http://dx.doi.org/10.1038/s41746-021-00544-y.
30. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. BMJ Open. 2021;11:e047709, http://dx.doi.org/10.1136/bmjopen-2020-047709.
31. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell. 2020;2:e200029, http://dx.doi.org/10.1148/ryai.2020200029.
32. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34, http://dx.doi.org/10.1016/j.jclinepi.2015.06.005.
33. Pieper D, Puljak L, González-Lorenzo M, Minozzi S. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J Clin Epidemiol. 2019;108:26–33, http://dx.doi.org/10.1016/j.jclinepi.2018.12.004.