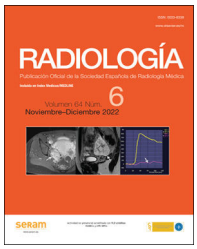
Radiología 66 (2024) 326–339

RADIOLOGÍA

www.elsevier.es/rx



ORIGINAL

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

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

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

Recibido el 6 de diciembre de 2022; aceptado el 18 de enero de 2023
Disponible en Internet el 20 de febrero de 2023

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

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 enfermedad. 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 enfermedad 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 favor de la capacidad diagnóstica de la intervención. Todas las RS obtuvieron la calificación «críticamente baja» siguiendo criterios AMSTAR-2.

* Autor para correspondencia.
Correo electrónico: jaime.vidal@salud.madrid.org (J. Vidal-Mondéjar).

https://doi.org/10.1016/j.rx.2023.01.007

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.

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

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

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 ''critically low'' following AMSTAR-2 criteria.
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.

Introducción

La radiografía de tórax es una técnica ampliamente utilizada en la práctica clínica. Presenta mayor facilidad de realización, menor coste y dosis de radiación con respecto a otras técnicas más complejas, como la tomografía computarizada (TC)[1,2]. Se estima que mundialmente, sólo entre 1997 y 2007, se realizan unos 3.600 millones de pruebas diagnósticas, de las cuales un 40% corresponde a radiografías de tórax[3]. La radiografía de tórax es interpretada por una gran variedad de profesionales médicos y requiere de entrenamiento para su adecuada lectura[2]. Uno de los retos actuales en los sistemas sanitarios es el aumento de la demanda de pruebas diagnósticas acompañado de una escasez de especialistas para su lectura, por lo que no es infrecuente que los servicios de radiodiagnóstico tengan dificultades para leer e informar todas las pruebas realizadas[4,5].

Entre los avances que podrían suponer importantes cambios en el diagnóstico por imagen destaca el uso de la inteligencia artificial (IA)[6]. La IA es un campo de la informática que pretende imitar la inteligencia humana con sistemas informáticos a través de la comparación iterativa de patrones complejos, generalmente a una velocidad y escala que superan la capacidad humana[7]. La IA podría tener múltiples aplicaciones dentro del campo de la imagen médica, como el diagnóstico asistido por ordenador, la detección de estudios de baja calidad o la selección automática de parámetros técnicos de los sistemas de obtención de imagen, entre otras[8–11]. Los sistemas de IA necesitan un importante volumen de pruebas de imagen para realizar su entrenamiento y alcanzar una adecuada precisión. En los últimos años se está estudiando el número de pruebas necesarias para que cada algoritmo funcione correctamente, que es variable en función del tipo de estudio[11]. Por otro lado, ha habido un creciente volumen de trabajos publicados al respecto, pero muchos de estos estudios individuales podrían presentar una calidad deficiente a nivel metodológico, así como en la presentación de los datos[12].

Las revisiones sistemáticas (RS) pretenden reunir toda la evidencia empírica, con el fin de responder una pregunta

específica de investigación. Utilizan métodos sistemáticos y explícitos, que se establecen con el fin de minimizar sesgos, aportando así resultados más fiables a partir de los cuales se puedan extraer conclusiones[13,14]. Las RS realizadas de forma adecuada, y siguiendo las guías o estándares metodológicos, son herramientas que proporcionan un elevado nivel de evidencia. Se han publicado algunas RS relacionadas con los avances en las técnicas diagnósticas que incorporan la IA[12,15–18]. Sin embargo, hasta la fecha no hay publicados trabajos que evalúen la calidad metodológica de las RS publicadas de estudios sobre diagnóstico con radiografía de tórax que utilizan sistemas de IA. El objetivo principal de este estudio fue evaluar la calidad metodológica de aquellas RS que utilizan técnicas basadas en IA para el diagnóstico mediante radiografía simple de tórax.

Material y métodos

Se trata de una RS metodológica de revisiones de pruebas diagnósticas.

Diseño del estudio y registro del protocolo

Para llevar a cabo este estudio se elaboró un protocolo que fue registrado prospectivamente (*Open Science Framework*: https://osf.io/4b6u2/). Durante la realización de este estudio no hubo variaciones significativas con respecto a los objetivos preestablecidos. La presentación de esta RS metodológica se realizó siguiendo las recomendaciones de la declaración PRISMA 2020 (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*)[19] (pp. 3-5 en el material suplementario).

Criterios de elegibilidad

Se seleccionaron los artículos atendiendo a criterios relacionados con el diseño/tipo de estudio, la población, las pruebas realizadas, los resultados, el tipo de publicación y el idioma.

Tipos de estudio: se incluyeron RS con y sin metaanálisis de cualquier tipo de estudio (por ejemplo, ensayos clínicos y estudios observacionales). Se consideraron aquellas RS que establecían explícitamente métodos para identificar estudios (por ejemplo, una estrategia de búsqueda), métodos explícitos de selección de estudios (por ejemplo, criterios de elegibilidad) y métodos descritos de síntesis (con o sin datos cuantitativos).

Población: las RS debían incluir estudios realizados en personas sanas y/o con cualquier indicio de enfermedad torácica sometidas a una radiografía de tórax, sin excluir por etnia, sexo ni edad.

Intervención y comparador: las RS debían evaluar el uso de algún sistema de IA, tanto sistemas de diagnóstico asistido por ordenador como de «aprendizaje profundo» (en inglés *deep learning*), «aprendizaje automático o de máquina» (en inglés *machine learning*) para la interpretación diagnóstica de la radiografía simple de tórax. Se incluyeron tanto radiografías convencionales como radiografías digitales. Se incluyeron aquellas RS cuyo propósito principal fuera la evaluación de la radiografía de tórax. Aquellas RS de múltiples modalidades diagnósticas (por ejemplo, TC, ultrasonidos...) donde se evaluaran sistemas de IA, a pesar de que incluyeran la radiografía simple de tórax, fueron excluidas. Como comparador o prueba de referencia debía presentarse al menos la práctica habitual o estándar de manejo clínico (por ejemplo, la interpretación por parte de un radiólogo, o confirmación microbiológica, en el caso de enfermedades infecciosas).

Resultados de interés: las RS debían presentar cualquier interpretación de la radiografía de tórax. Los estudios incluidos en las RS debían evaluar la precisión o fiabilidad diagnóstica de las pruebas (por ejemplo, en términos de sensibilidad, especificidad y valores predictivos).

Duración del seguimiento: no se estableció ningún límite en cuanto a la duración del seguimiento.

Estado de la publicación: se evaluaron tanto RS publicadas como en estado de espera de publicación (prepublicación).

Idiomas: se incluyeron RS escritas en inglés o español.

Fuentes de información y estrategia de búsqueda

Para identificar los artículos se realizó una búsqueda exhaustiva en las principales bases de datos (desde sus inicios hasta el 26 de mayo de 2022): MEDLINE (a través de PubMed), EMBASE, y Cochrane Database of Systematic Reviews. Con ayuda de una experta documentalista se diseñaron las estrategias de búsqueda (CA-FM, Biblioteca Nacional de Ciencias de la Salud, Instituto de Salud Carlos III) que incluyeron palabras clave relacionadas con la IA, la radiografía de tórax y RS/metaanálisis (pp. 6 y 7 del material suplementario). Adicionalmente, se examinaron Google Académico y las referencias bibliográficas de los artículos potencialmente seleccionables, contactando con los autores en caso de necesitar información adicional, con el propósito de aumentar la sensibilidad de las búsquedas.

Selección de los estudios

Dos investigadores (J. V-M y L. T-R) realizaron la selección de las RS siguiendo los criterios de inclusión y exclusión; las discrepancias entre ellos fueron discutidas con un tercer investigador (F. C-L) para llegar a un acuerdo sobre ellas. Para ello se utilizó el software Rayyan® (Rayyan Systems Inc., Cambridge, EE. UU.)[20], procediéndose a la eliminación de los artículos duplicados.

Recogida de información

Dos investigadores (J. V-M y L. T-R) realizaron la extracción de datos relevantes de las RS incluidas de manera independiente, como son:

Características generales de las RS: primer autor y año de publicación, país del autor de correspondencia, nombre de la revista y factor de impacto (de acuerdo con la Journal *Citation Reports*; 2021), número de bases de datos utilizadas y nombres (por ejemplo, PubMed, EMBASE, Scopus y otras), mención de existencia de un protocolo (sí/no), y en caso afirmativo, dónde está accesible (por ejemplo, PROSPERO), descripción de las herramientas utilizadas, tanto

para su presentación (por ejemplo, declaración PRISMA[19]) como para la evaluación de la calidad metodológica (por ejemplo, herramienta *A Measurament Tool to Assess Systematic Reviews* 2[21,22] [AMSTAR-2]), uso adecuado/inadecuado de PRISMA (de acuerdo con criterios definidos por Caulley et al.)[23] y presentación de una lista de comprobación cumplimentada, mención de algún instrumento para evaluar el riesgo de sesgo de los estudios (por ejemplo, *Quality Assessment of Diagnostic Accuracy Studies* [QUADAS-2]) y fuente de financiación.

Características específicas de las RS: descripción de la intervención (por ejemplo, nombre del algoritmo evaluado, tipo de arquitectura, fuentes de datos de entrenamiento y número de imágenes), descripción del comparador (por ejemplo, confirmación diagnóstica como práctica habitual y pruebas microbiológicas), número y diseño de los estudios incluidos (por ejemplo, ensayos aleatorizados y estudios observacionales), descripción de las características de los participantes en los estudios incluidos, si se realizó alguna validación y de qué tipo, medida de precisión utilizada (por ejemplo, sensibilidad y especificidad).

Métodos de síntesis empleados en las RS: tipo de síntesis (por ejemplo, narrativa/cualitativa y cuantitativa/metaanálisis), modelo de análisis utilizado si procede, presentación del efecto combinado y su intervalo de confianza del 95% y realización de análisis adicionales (por ejemplo, análisis de subgrupos y metarregresión).

Resultados o conclusiones cualitativas presentadas en las RS: favorables si el sistema de IA evaluado era claramente la opción recomendada (por ejemplo, se menciona como «eficaz», «beneficioso», «mejora la precisión diagnóstica», «técnica prometedora»), desfavorables si las conclusiones eran claramente negativas (por ejemplo, «no es eficaz», «es poco probable que sea beneficiosa», «no mejora el diagnóstico») y neutrales o no concluyentes cuando la intervención de interés no era superior al comparador o cuando las conclusiones se expresaron con un elevado grado de incertidumbre.

Evaluación de la transparencia y la calidad metodológica

Dos investigadores (J. V-M y L. T-R) evaluaron la transparencia y la calidad metodológica de las RS incluidas. Para evaluar la calidad metodológica se utilizó AMSTAR-2[21,22]. AMSTAR-2 presenta una lista de comprobación que presenta opciones de respuesta breve («sí», «sí parcial» y «no») que evalúan distintos dominios relevantes de la RS[21,22]. Estos dominios se subdividen en 7 «críticos» y 9 «no críticos» en función de si se considera que podría afectar de forma significativa a la validez de los resultados. Los resultados de cada ítem pueden obtener una respuesta de «sí» en el caso de cumplirse de forma completa, «sí parcial» bajo ciertos supuestos variables en cada ítem o con cumplimiento parcial y «no» en caso de no cumplir los criterios del ítem. Se establecen 4 opciones de calificación o confianza: «alta» (ausencia de debilidades críticas y máximo de una debilidad no crítica), «moderada» (ausencia de debilidades críticas y 2 o más debilidades no críticas), «baja» (máximo de una debilidad crítica independientemente del número de debilidades no críticas) y «críticamente baja» (2 o más debilidades críticas)[21,22].

La herramienta utilizada para evaluar la transparencia en la presentación de las RS incluidas fue la extensión de la declaración PRISMA para RS de pruebas diagnósticas (PRISMA-DTA)[24]. La declaración PRISMA-DTA[24] fue publicada en enero de 2018, e incluye una lista de comprobación con 27 ítems diseñada para mejorar la integridad del informe de los métodos y resultados de RS de pruebas diagnósticas.

Análisis de datos

Se realizó una síntesis narrativa de las características principales de las RS incluidas, y todos los datos extraídos se expusieron en tablas de evidencia. Para cada RS se presentaron elementos generales, metodológicos y de transparencia. Se realizó análisis descriptivo, mediante recuento de frecuencias y porcentajes, de los resultados obtenidos en las listas de comprobación AMSTAR-2[21,22] y PRISMA-DTA[24].

Resultados

Resultados de las búsquedas y selección de las revisiones sistemáticas incluidas

A partir de las búsquedas se identificaron 797 registros, eliminando 35 duplicados. De ellos se excluyeron 733 por considerarse irrelevantes tras la lectura del título y el resumen. Finalmente se evaluaron 29 artículos mediante lectura a texto completo. Tras excluir 22 artículos (ver pp. 8 y 9 del material suplementario), se incluyeron 7 RS[16–18,25–28] (fig. 1).

Características generales de las revisiones sistemáticas

Las características generales de las RS incluidas se presentan en la tabla 1. Las 7 RS mencionan las bases de datos para las búsquedas, y todas consultaron al menos 3 (rango: 3-5 bases datos). Solo 2 (29%) RS presentan un protocolo descrito y accesible. Seis (86%) RS mencionan utilizar PRISMA o alguna de sus extensiones, aunque solo una de ellas incluyó la lista de comprobación cumplimentada (por ejemplo, en anexos). La evaluación del riesgo de sesgos fue realizada en 4 (57%) RS, utilizando principalmente la herramienta QUADAS-2 (n = 3; 43%).

Características específicas de las revisiones sistemáticas

En la tabla 2 se describen las características específicas tanto de las RS como de los estudios incluidos. Las 7 RS evaluaron sistemas de IA basados en técnicas de «aprendizaje profundo». En 2 (29%) RS se mencionan además técnicas basadas en «aprendizaje automático». Las 7 RS evaluaron la capacidad diagnóstica de sistemas de IA dirigidos a enfermedades torácicas infecciosas (por ejemplo, tuberculosis: n = 4; 57%; neumonía: n = 2; 29%; y COVID-19: n = 1). Una RS[17] se basaba en lectura automática de la radiografía de tórax en población pediátrica. Solo 2 (29%) RS[26,27] definieron claramente cuántas imágenes fueron utilizadas (por ejemplo, tanto para el entrenamiento como para la evaluación de los sistemas de IA). Cinco (71%) RS realizaban una descripción

Tabla 1 Características generales de las RS incluidas

Autor y año de publicación	País del autor de correspondencia	Nombre de la revista	Factor de impacto (2020)	Bases de datos utilizadas, número (nombre)	Presencia de protocolo y accesibilidad del mismo	Herramientas de presentación metodológica	Lista de comprobación PRISMA cumplimentada	Instrumento de evaluación de riesgo de sesgo	Fuente de financiación
Ghaderzadeh et al.[25], 2021	Irán	Biomed Res Int	3,411	4 (Scopus, Elsevier ScienceDirect, PubMed y Web of Science)	No	PRISMA 2009	No	CHARM checklist modificada	Pública
Harris et al.[26], 2019	Canadá	PLoS One	3,240	4 (PubMed, MEDLINE, EMBASE y Scopus)	Sí, PROSPERO	PRISMA 2009	Sí	QUADAS-2	Ninguna
Li et al.[27], 2020	China	Comput Biol Med	4,589	5 (PubMed, Embase, Scopus, Web of Science y Google Scholar)	No	PRISMA 2009	No	No	Ninguna
Oloko-Oba et al.[28], 2022	Sudáfrica	Front Med	5,093	4 (Scopus, IEEE Xplore, Web of Science y PubMed)	No	PRISMA 2009	No	No	No descrita
Padash et al.[17], 2022	Canadá	Pediatr Radiol	2,505	3 (PubMed, EMBASE y Web of Science)	No	PRISMA-DTA	No	No	No descrita
Pande et al.[16], 2016	Canadá	Int J Tuberc Lung Dis	2,373	4 (PubMed, EMBASE, Scopus y Engineering Village)	No	No	No	QUADAS-2	Pública
Tavaziva et al.[18], 2021	Canadá	Clin Infect Dis	9,079	5 (MEDLINE, EMBASE, PubMed, Scopus e Engineering Village)	Sí, PROSPERO	PRISMA-IPD	No	QUADAS-2	Pública

CHARMS: *CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies*; IEEE: *Institute of Electrical and Electronics Engineers*; QUADAS-2: *Quality Assessment of Diagnostic Accuracy Studies 2*; 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*.

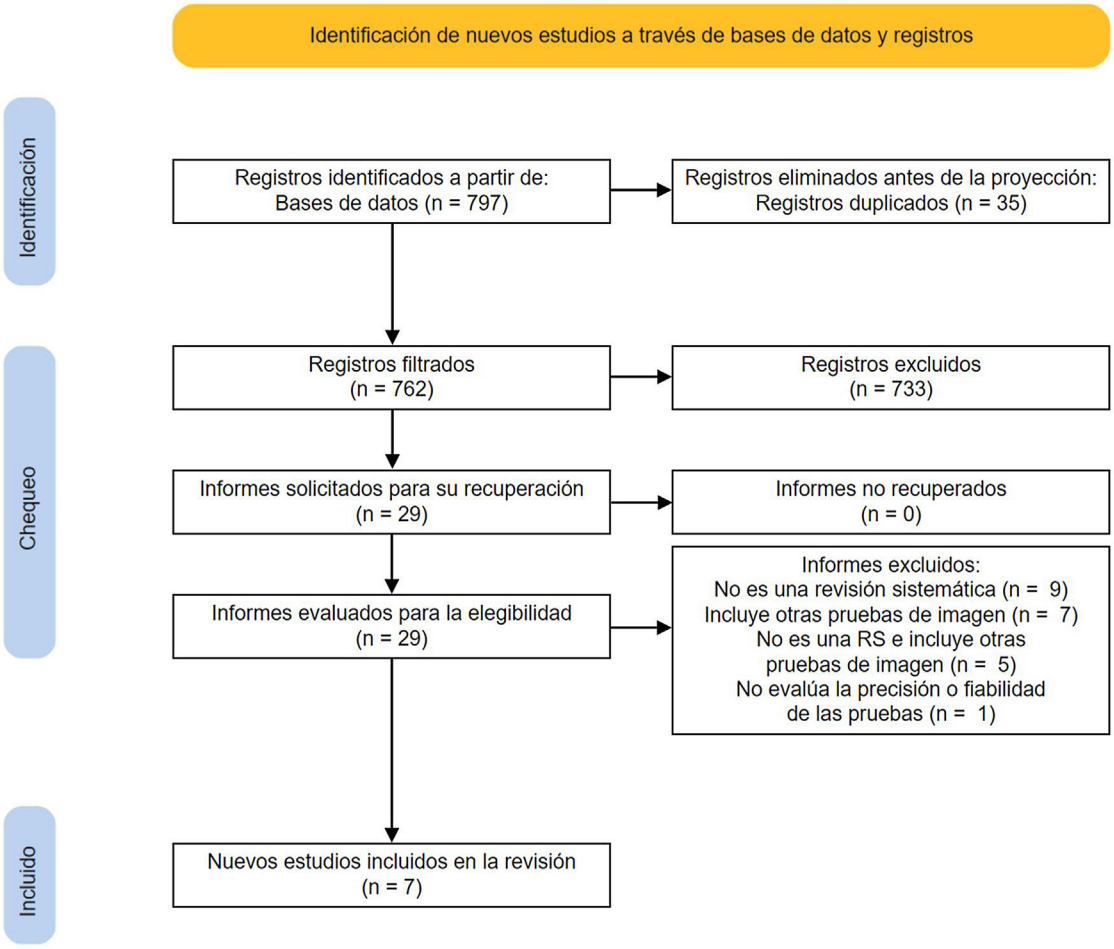


Figura 1 Diagrama de flujo para la selección de estudios.

de cuáles fueron las fuentes de información utilizadas para la obtención de las imágenes radiográficas. Cuatro (57%) RS describieron el comparador utilizado (por ejemplo, estándar de referencia microbiológico de la enfermedad infecciosa y/o criterios clínicos).

Las 7 RS presentaron el número de estudios incluidos (media = 36 estudios y rango: 4-62 estudios por revisión). Ninguna RS presentó una descripción del diseño de los estudios incluidos. Tres RS (43%) mencionaron algunas características generales de los participantes de los estudios. Las 7 RS describieron las medidas de precisión de las pruebas, como la sensibilidad, especificidad o área bajo la curva (ABC) (tabla 2). Solo una RS[17] presentó que en algunos de los estudios incluidos habían realizado alguna validación, tanto interna como externa de los sistemas de IA. En el resto de RS (n = 6; 86%) no hay ninguna referencia a una posible validación.

La descripción de los métodos de síntesis utilizados en las RS se presenta en la tabla 3. En 5 RS (71%) no se realizó metaanálisis. En una de las RS con metaanálisis[27] se realizaba una comparación de la capacidad diagnóstica de los sistemas de IA para el diagnóstico de neumonía y para la distinción entre neumonía viral y bacteriana (por ejemplo, sensibilidad = 0,98; especificidad = 0,94 y ABC de 0,99). En otra RS con metaanálisis y metarregresión[18] evaluaron distintos sistemas de IA para la evaluación de la tuberculosis pulmonar (por ejemplo, ABC = 0,83-0,85 en función de distintos sistemas software de IA; especificidad = 0,54-0,60).

Resultados o conclusiones cualitativas de las revisiones sistemáticas

Cinco (71%) RS presentaron resultados «favorables» a la utilización de la IA como un sistema que puede aportar precisión y fiabilidad para el diagnóstico de enfermedades infecciosas con radiografía de tórax. Tan solo 2 RS[16,18] fueron algo más críticas a la hora de presentar e interpretar estos resultados (tabla 4).

Resultados de la evaluación de la transparencia y la calidad metodológica

En la evaluación de la calidad metodológica, las 7 RS obtuvieron una calificación «críticamente baja» según los criterios AMSTAR-2 (ver pp. 10 y 11 del material suplementario). Por ejemplo, solo 2 RS[18,26] obtuvieron un resultado de «sí parcial» (referente a la presencia de un protocolo, descripción previa de los pasos que se van a seguir durante la revisión y la justificación de cualquier desviación

Tabla 2 Características específicas de las RS incluidas

Autor y año de publicación	Descripción de la intervención	Descripción del comparador	Número y diseño de estudios incluidos en la revisión	Descripción de las características de los participantes	Presencia de validación y tipo (interna/ externa)	Medida de precisión utilizada
Ghaderzadeh et al.[25], 2021	1. Métodos de IA basados en ML y DL en radiografía de tórax en pacientes con COVID-19 2. Diferencian los sistemas de IA en ''función de su objetivo (clasificar, detectar lesiones, segmentación...) 3. N° imágenes entrenamiento: No definido 4. N° imágenes testadas: No definido. 5. Descripción de las bases de datos utilizadas: Sí	No hacen descripción completa del comparador	60 estudios No especifica diseño	Se recogen las características generales de los pacientes: No Otros datos que describen a los participantes: Diferencian entre estudios dicotómicos (COVID sí/no) vs estudios con más variables N° participantes: 86 - 13975 Enfermedad estudiada: COVID-19	No/No	Sensibilidad, especificidad, precisión, ABC, valor F, exhaustividad
Harris et al.[26], 2019	1. Softwares basados en ML (n= 46) y DL (n=7) en radiografía de tórax en pacientes con tuberculosis. 2. Múltiples algoritmos basados en IA (CAD4TB, muchos otros sin definir...) 3. N° imágenes entrenamiento: 18-60989. 4. N° imágenes testadas: 30- 37475. 5. Descripción de las bases de datos utilizadas: Sí	Estándar de referencia microbiológico o inicio de tratamiento: 17 Lector humano: 39	53 estudios No especifica diseño	Se recogen las características generale de los pacientes: Sí Otros datos que describen a los participantes: Estudios de triaje vs estudios de screening N° participantes: 161 - 17006 Enfermedad estudiada: tuberculosis	No/No	ABC, sensibilidad, especificidad, verdaderos positivos, falsos positivos, ratio falsos positivos, verdaderos negativos, falsos negativos
Li et al.[27], 2020	1. Métodos de IA basados en DL. 2. Diagnóstico de enfermedad vs. normalidad y diferenciación entre tipos de neumonías (viral y bacteriana). 3. N° imágenes entrenamiento: 2000 − 25648. 4. N° imágenes testadas: 300 − 1928 5. Descripción de las bases de datos utilizadas: Sí	No se hace descripción completa del comparador	15 estudios No especifica diseño	Se recogen las características generales de los pacientes: No Otros datos que describen a los participantes: No N° participantes: No definido Enfermedad estudiada: neumonía, viral y bacteriana	No/No	ABC, sensibilidad, especificidad, verdaderos positivos, falsos positivos, razón de verosimilitud negativa, razón de verosimilitud positiva, verdaderos negativos, falsos negativos, odds ratio diagnóstico

Tabla 2 (*continuación*)

Autor y año de publicación	Descripción de la intervención	Descripción del comparador	Número y diseño de estudios incluidos en la revisión	Descripción de las características de los participantes	Presencia de validación y tipo (interna/externa)	Medida de precisión utilizada
Oloko-Oba et al.[28], 2022	1. Métodos de IA basados en DL. 2. Técnicas de DL para el diagnóstico de tuberculosis 3. N° imágenes entrenamiento: No definido 4. N° imágenes testadas: No definido 5. Descripción de las bases de datos utilizadas: Sí	No se hace descripción completa del comparador, únicamente en uno de los estudios incluidos se indica que el comparador son los médicos	62 estudios No especifica diseño	Se recogen las características generales de los pacientes: No Otros datos que describen a los participantes: No definido N° participantes: No definido Enfermedad estudiada: tuberculosis	No/No	ABC, sensibilidad, especificidad, precisión, exhaustividad, valor F
Padash et al.[17], 2022	1. Descripción de las bases de datos de radiografía pediátrica 2. Uso del DL para el diagnóstico de neumonía y otras enfermedades 3. N° imágenes entrenamiento: No definido 4. N° imágenes testadas: No definido 5. Descripción de las bases de datos utilizadas: Sí	Evaluación radiológica o clínica	55 estudios No especifica diseño	Se recogen las características generales de los pacientes: Sí Otros datos que describen a los participantes: Grupos de edad. N.° participantes: No Enfermedad estudiada: Neumonía (mayoría de los estudios), FQ, SDR, bronquitis/bronquiolitis, neumotórax	Hay validación interna y externa en algunos de los trabajos	ABC, sensibilidad, especificidad, precisión
Pande et al.[16], 2016	1. Métodos de IA basados en DL en radiografía. 2. Algoritmo de DL (CAD4TB) para el diagnóstico de tuberculosis 3. N° imágenes entrenamiento: No definido 4. N° imágenes testadas: No definido 5. Descripción de las bases de datos utilizadas: No	Estándar de referencia microbioló-gico (esputo, cultivo o PCR) (5) Cultivo + criterios clínicos (1)	5 estudios No especifica diseño	Se recogen las características generales de los pacientes: Parcialmente Otros datos que describen a los participantes: País, % VIH N° participantes: 161 - 894 Enfermedad estudiada: Tuberculosis	No/No	ABC, Sensibilidad, Especificidad

Tabla 2 (*continuación*)

Autor y año de publicación	Descripción de la intervención	Descripción del comparador	Número y diseño de estudios incluidos en la revisión	Descripción de las características de los participantes	Presencia de validación y tipo (interna/externa)	Medida de precisión utilizada
Tavaziva et al.[18], 2021	1. Métodos de IA basados en DL en radiografía 2. Uso de varios algoritmos de DL para el diagnóstico de tuberculosis 3. N° imágenes entrenamiento: No definido 4. N° imágenes testadas: No definido 5. Descripción de las bases de datos utilizadas: No	Estándar de referencia microbiológico (Esputo, cultivo o PCR) En un análisis posterior, comparación con lector humano	4 estudios No especifica diseño	Se recogen las características generales de los pacientes: Sí Otros datos que describen a los participantes: Estatus VIH, resultado cultivo... N° participantes: 342 - 2298 Enfermedad estudiada: Tuberculosis	No/No	ABC, sensibilidad, especificidad, razón de verosimilitud positiva, razón de verosimilitud negativa

ABC: área bajo la curva; DL: *deep learning*; IA: inteligencia artificial; FQ: fibrosis quística; ML: *machine learning*; PCR: *polymerase chain reaction*; SDR: Síndrome de distrés respiratorio; VIH: virus de la inmunodeficiencia humana.

significativa del protocolo). Ninguna RS cumplió el dominio crítico 7 (presentación de una lista de los estudios excluidos con las razones de dicha exclusión).

La transparencia y calidad de la presentación de las RS incluidas fue variable según los criterios de la declaración PRISMA[19,24]. Por ejemplo, 4 RS (57%) no describieron adecuadamente la estrategia de búsqueda (ítem 8), 3 (43%) no describieron la evaluación del riesgo de sesgos (ítem 13), y 4 (57%) no presentaron adecuadamente cómo se realizó la síntesis de los resultados (ítem 14). Solo una RS[16] describió adecuadamente las características de los estudios. Solo 2 RS presentaron adecuadamente la valoración de riesgo de sesgo obtenido (ítem 19). En el material suplementario (pp. 12 y 13) se presenta la cumplimentación detallada de la declaración PRISMA-DTA[24] para cada RS. Cuando se examinó el uso apropiado de PRISMA por parte de los autores en las 6 RS que las citan, tan solo 2 RS[17,27] parecen utilizarla de manera adecuada. En el resto de RS se observó un uso inadecuado (en una RS[16]) o quedan dudas (en 3 RS[25,26,28]).

Discusión

El principal hallazgo de esta investigación ha sido demostrar que la calidad metodológica de las RS publicadas que evalúan el uso de IA en radiografía de tórax es muy variable, existiendo una falta de cumplimiento en muchos de los ítems propuestos en las guías y estándares metodológicos utilizados. Las herramientas metodológicas como AMSTAR-2[21,22] permiten identificar elementos que influyen en la calidad del proceso de realización de una revisión, aspecto fundamental para interpretar y evaluar la potencial aplicabilidad de sus resultados.

Para que una RS pueda considerarse de «alta» calidad metodológica (según los criterios AMSTAR-2[21,22]) se debería obtener un resultado favorable en la práctica totalidad de los aspectos evaluados. Sin embargo, en las RS analizadas en este estudio se identifican varios dominios críticos que no se cumplen. Por ejemplo, uno de los aspectos más importantes del diseño de una RS es la elaboración y registro de un protocolo donde se establezcan (previa a la realización) los métodos que se van a seguir para su desarrollo. En este sentido, el dominio que evalúa AMSTAR-2[21,22] hace referencia concretamente a estos aspectos, y resulta destacable que solo 2 RS presentaron una valoración de «sí parcial», aspecto que puede comprometer significativamente la calidad metodológica de estos trabajos. Otro aspecto fundamental en la realización de una RS es la inclusión de un listado de los estudios excluidos (con la justificación o motivos de dichas exclusiones). La ausencia de esta información en las RS analizadas podría comprometer la calidad de una RS al no poder descartar posibles sesgos de selección de los estudios. Existen otros aspectos cuya interpretación puede ser más compleja a la hora de evaluar la calidad metodológica. Por ejemplo, en el caso de AMSTAR-2[21,22] se precisa conocer la definición de los tipos de estudio incluidos en la revisión. Sin embargo, en tan solo una de las RS analizadas se hizo una descripción detallada y clara sobre cuál era el tipo de estudios incluidos. Una razón que podría explicar la ausencia de estas definiciones es que al menos una parte de los investigadores considerara algo evidente que se trata de estudios de evaluación diagnóstica. Sería necesario que futuras revisiones presentaran la descripción detallada y pormenorizada previa y durante la realización de la revisión del tipo de estudios (por ejemplo, experimentales u observacionales).

A diferencia de AMSTAR-2[21,22], las guías de publicación, como la declaración PRISMA[19] y su extensión PRISMA-DTA[24], tienen como objetivo principal la ayuda en la presentación completa y transparente de los distintos apartados de una RS. Sin embargo, que los autores citen la declaración

Tabla 3 Métodos de síntesis aplicados en las revisiones sistemáticas incluidas

Primer autor y año de publicación	Tipo de síntesis	Modelo de análisis utilizado	Presentación del efecto combinado del metaanálisis	Realización de análisis adicionales
Ghaderzadeh et al.[25], 2021	Cualitativa (narrativa) y cuantitativa (sin metaanálisis)	No procede	No procede	Comparan sensibilidad y especificidad con otros estudios
Harris et al.[26], 2019	Cualitativa (narrativa) y cuantitativa (sin metaanálisis)	No procede	No procede	ABC según estudio (desarrollo/clínico) y por tipo (DL/ML)
Li et al.[27], 2020	Cualitativa (narrativa) y cuantitativa (con metaanálisis)	No se describe	Sensibilidad: 0,98 (95% IC: 0,96-0,99), Especificidad: 0,94 (95% IC: 0,90-0,96) DOR: 718,13 (95% IC: 288,45-1787,93) ABC de curva SROC: 0,99 (95% IC: 0,98-100) LHR+: 96,1 (95% IC: 96,1-97,7) LHR−: 0,02 (95% IC: 0,01-0,04)	Analizan los estudios con datos de neumonía viral y bacteriana; Sensibilidad: 0,89 (95% IC: 0,79-0,94) Especificidad: 0,89 (95% IC: 0,78-0,95) DOR: 66,14 (95% IC: 17,34-252,37) ABC de curva SROC: 0,95 (0,93-0,97) RV+: 8,34 (95% IC: 3,75-18,55) RV−: 0,13 (95% IC: 0,06-0,26)
Oloko-Oba et al.[28], 2022	Solo cualitativa (narrativa)	No procede	No procede	No se realizaron
Padash et al.[17], 2022	Solo cualitativa (narrativa)	No procede	No procede	No se realizaron
Pande et al.[16], 2016	Solo cualitativa (narrativa)	No procede	No procede	No se realizaron
Tavaziva et al.[18], 2021	Cualitativa (narrativa) y cuantitativa (con metaanálisis de datos individuales de pacientes)	Modelo de efectos aleatorios	ABC: CAD4TBv6, 0,83 (95% IC: 0,82-0,84); Lunit, 0,83 (95% IC: 0,79-0,86); qxRv2, 0,85 (95% IC: 0,83-0,88) Especificidad con un 90% de sensibilidad: CAD4TBv6, 0,57 (95% IC0,52-0,62); Lunit, 0,54 (95% IC: 0,45-0,63; qXRv2, 0,60 (95% IC: 0,52-0,69)	Análisis de subgrupos (sexo, VIH, esputo, antecedentes de tuberculosis y edad); Metarregresión; análisis *post-hoc* comparando con lectores humanos

ABC: área bajo la curva; DL: *deep learning*; DOR: diagnostic *odds ratio*; IC: intervalo de confianza; ML: *machine learning*; RV+: razón de verosimilitud positiva; RV−: razón de verosimilitud negativa; SROC: *summary receiver operating characteristic*; VIH: virus de la inmunodeficiencia humana.

PRISMA[19] no garantiza la correcta utilización de la guía, ni tampoco que la RS se haya realizado de una manera rigurosa y transparente. En este sentido, cuando se examinó el uso de la declaración PRISMA por parte de los autores de las RS incluidas, de las 6 RS que la citan y manifiestan su uso, tan solo 2 de ellas parecen emplearla de forma adecuada. Estos resultados estarían en la línea de otros trabajos recientes[23], donde se identificó que un número importante de artículos de investigación no utiliza de manera adecuada este tipo de guías. En este sentido, parecería necesario promover la formación sobre las principales guías de publicación (como la declaración PRISMA[19,24]), tanto para los autores de las RS, como para los revisores y los editores de las revistas en las que se publican.

El cumplimiento de los distintos apartados de las guías y estándares metodológicos es fundamental para facilitar la lectura y la comprensión de cualquier investigación, así como para poder evaluar de forma rápida y crítica la

Tabla 4 Presentación de los resultados cualitativos en las revisiones sistemáticas incluidas

Primer autor y año de publicación	Resultado/conclusión por parte de los autores de la RS sobre el uso de la IA	Resultado cualitativo
Ghaderzadeh et al.[25] 2021	Las radiografías equipadas con IA pueden utilizarse como una herramienta de *screening* de algunos casos que requieran TC. El uso de esta herramienta no consume tiempo o supone costes extra, tiene mínimas complicaciones y puede disminuir o eliminar TC innecesarias u otros recursos del sistema de salud	A favor de la intervención
Harris et al.[26], 2019	Concluimos que los programas de diagnóstico asistido por ordenador son prometedores pero la mayoría del trabajo se ha centrado en su desarrollo más que en su evaluación clínica. Aportamos sugerencias concretas sobre qué elementos del diseño de los estudios deben mejorarse	A favor de la intervención
Li et al.[27], 2020	El aprendizaje profundo *(deep learning)* presenta un buen rendimiento diagnóstico para diferenciar neumonía de una radiografía normal, así como en distinguir neumonía viral y bacteriana. Sin embargo, algunas preocupaciones metodológicas importantes deben tenerse en cuenta en estudios futuros para trasladarlo a la clínica	A favor de la intervención
Oloko-Oba et al.[28], 2022	Concluimos que los sistemas de diagnóstico asistido por ordenador son prometedores para el abordaje de los retos de la epidemia de tuberculosis y hacemos recomendaciones para la mejora en estudios futuros	A favor de la intervención
Padash et al.[17], 2022	La clasificación de las radiografías de tórax como neumonía es la principal aplicación de la IA, evaluada en el 65% de los estudios. A pesar de que la mayoría de los estudios reportan una alta precisión diagnóstica, muchos algoritmos no están validados en bases de datos externas. La mayoría de los estudios de IA para la interpretación de la radiografía de tórax en población pediátrica se centran en un limitado número de enfermedades, y su progreso se encuentra obstaculizado por la falta de bases de datos de gran tamaño de radiografías de tórax en población pediátrica	Neutral
Pande et al.[16], 2016	La evidencia evaluando la precisión diagnóstica del diagnóstico asistido por ordenador está limitada por el escaso número de estudios, muchos de los cuales presentan importantes limitaciones metodológicas, la disponibilidad y evaluación de solo un programa de software y la generabilidad limitada de los ajustes donde la tuberculosis o el VIH son menos prevalentes. Se requiere de investigación adicional	Neutral
Tavaziva et al.[18], 2021	Para el análisis de la implementación del diagnóstico asistido por ordenador de radiografía de tórax como un test de alta sensibilidad para descartar la tuberculosis, los usuarios necesitan un punto de corte identificado de sus propias poblaciones y estratificado por el estado de VIH y del esputo	A favor de la intervención

IA: inteligencia artificial; TC: tomografía computarizada.

presencia de esta información. En las RS incluidas destaca la amplia variabilidad a la hora de presentar la metodología empleada, así como la ausencia de presentación *a priori* de algunas de las medidas de precisión que van a analizarse o cómo se van a sintetizar los resultados de los estudios incluidos. También cabe destacar, en el apartado de resultados, la ausencia de datos de las características de los estudios, la escasa presentación de la evaluación del riesgo de sesgos o de los resultados de los estudios individuales. Debido a que es una herramienta que pretende aclarar la presentación de los métodos y resultados de las RS, la ausencia de muchos de estos apartados limita claramente la transparencia de las RS publicadas.

Diversos retos pueden plantearse a la hora de diseñar este tipo estudios, analizar los datos e interpretar los resultados de las investigaciones que evalúan la IA en imagen médica[16]. La mayoría de las RS analizadas parecen presentar resultados favorables al uso de sistemas de IA para la mejora de la capacidad diagnóstica de enfermedades infecciosas en radiografía de tórax. En algunos casos se añaden limitaciones sobre los resultados, aunque mayoritariamente los estudios parecen no proporcionar una interpretación prudente de los resultados, al menos en función de la evaluación cualitativa de los resultados obtenidos a favor de los sistemas de IA. Cuando la presentación de parámetros de precisión diagnóstica se realiza sin hacer una descripción detallada y completa de cuál es el comparador puede dificultar las interpretaciones de los resultados de los estudios primarios, pudiendo dar lugar a errores. Por ejemplo, al realizar

la comparación de la precisión diagnóstica de un software de IA con un lector humano (no necesariamente informado por un radiólogo), lo que se compara es la capacidad de la lectura de la radiografía, mientras que, al compararlo con parámetros de la enfermedad como las referencias microbiológicas, se compara la precisión para el diagnóstico de una enfermedad. En este sentido, en la práctica totalidad de RS incluidas en este trabajo no había una definición clara del comparador, por lo que consideramos que los resultados son difícilmente interpretables y extrapolables a un escenario clínico real. Podría pensarse que muchas de las radiografías de tórax de las 7 RS no fueron informadas por radiólogos con una dedicación exclusiva (o preferente) a la enfermedad pulmonar. En el caso de que muchas no hubieran sido informadas por radiólogos torácicos, esto podría suponer una potencial fuente de error, ya que estaría demostrado que la inclusión de radiólogos mejora la calidad de las interpretaciones de las imágenes (tanto por una mayor experiencia como por disponer, generalmente, de visores diagnósticos y pantallas con mejor resolución espacial que otros médicos no radiólogos), pudiendo sobrevalorar los resultados de una herramienta de IA. Asimismo, en el caso de las radiografías de tórax (a diferencia de otras modalidades como la TC o la resonancia magnética [RM], que suelen ser interpretadas generalmente por radiólogos) la falta de definición clara del lector humano (radiólogo vs. médico no radiólogo) podría tener un enorme impacto en la interpretación (y extrapolación) de los resultados. Tal vez esta falta de definición del comparador sea una de las causas por las que en otros trabajos recientes[29] las modalidades «no axiales» (como la radiografía de tórax o los ultrasonidos) presentaban un riesgo significativamente mayor de sesgo en el dominio de «estándar de referencia» que las modalidades «axiales» (TC y RM).

La presentación de la información relativa a estos aspectos y otras características (como pueden ser las bases de datos de radiografías, los procedimientos de obtención de imágenes para el entrenamiento, su validación o el comparador) es mejorable en gran parte de los estudios incluidos en las RS analizadas. Se están desarrollando algunas propuestas o iniciativas como la extensión de la declaración *Standards for Reporting of Diagnostic Accuracy Studies* (STARD) para estudios de pruebas diagnósticas que aplican IA (STARD-AI)[30], además de la reciente publicación de la lista de comprobación *Checklist for Artificial Intelligence in Medical Imaging*[31], que podrían contribuir a mejorar la transparencia y calidad de este tipo de estudios. Aún no se dispone de herramientas específicas para la evaluación del riesgo de sesgos en RS de trabajos de IA, por lo que sería interesante el desarrollo y la aplicación de estas herramientas en futuras investigaciones. Además, futuros trabajos también debieran considerar otros aspectos éticos y legales de aplicación a los sistemas de IA. Por ejemplo, en virtud de la legislación vigente, el Reglamento General de Protección de Datos (RGPD) otorga a los pacientes europeos el derecho a la explicación de todas las decisiones tomadas por un algoritmo. Los futuros estudios que quieran demostrar la utilidad de los sistemas de IA tendrán que ser capaces de poder dar una explicación de cómo han conseguido esos resultados (transparencia) y también deberán obtener un consentimiento informado prospectivo de los pacientes para el uso y explotación de sus imágenes médicas. Por otro lado, en varios estudios no se hace evaluación del riesgo de sesgos y, en los casos donde sí que existe, se aporta muy poca información sobre dicha evaluación, tanto del propio estudio como de las bases de datos incluidas. En algunas de las RS incluidas se han realizado comparaciones de algunos resultados de la revisión con los datos obtenidos de otros trabajos para obtener alguna de las conclusiones del trabajo. Esto supondría un problema metodológico debido al alto riesgo de sesgos al utilizar datos que han seguido metodologías distintas para obtener resultados y que no forman parte del propio trabajo.

La presente RS metodológica se ha realizado siguiendo un protocolo creado *a priori*, sin que hayan ocurrido desviaciones significativas sobre el mismo. Se ha intentado ser transparentes a la hora de presentar tanto la selección de estudios como la extracción de información. Además, la evaluación de los principales aspectos de calidad metodológica se ha realizado por duplicado, y resolviendo las posibles discrepancias entre los investigadores, siguiendo las principales guías y estándares metodológicos (PRISMA[19,24] y AMSTAR-2[21,22]). Sin embargo, cabe destacar algunas limitaciones en este trabajo. En primer lugar, esta evaluación se ha realizado sobre las RS de una modalidad diagnóstica específica como es la radiografía de tórax, por lo que podría no ser aplicable en otros campos de la IA aplicada en imagen médica. Por otro lado, hay que señalar que la herramienta que aporta más información sobre la calidad metodológica de las RS (AMSTAR-2[21,22]) puede parecer muy exigente, al observar que todas las RS presentan deficiencias críticas. No se puede descartar que existan diferencias en los resultados si se hubieran aplicado otras herramientas para evaluar la calidad metodológica y riesgo de sesgo de las RS (por ejemplo, la escala ROBIS[32]), aunque de haberlas serían mínimas[33]. Por último, este trabajo evaluó la literatura publicada en revistas biomédicas indexadas en las principales bases de datos, por lo que no puede descartarse la falta u omisión de otras RS no publicadas, así como la potencial exclusión de artículos publicados en otros idiomas (en particular, del chino).

Conclusiones

Los resultados de este trabajo indican que la calidad metodológica de las RS que utilizan sistemas de IA en radiografía de tórax es mejorable, presentando deficiencias importantes muchas de ellas. La falta de cumplimiento de muchos de los elementos o características analizadas hace que las RS publicadas en este campo deban interpretarse con cautela. Aumentar la calidad metodológica y la transparencia de estas RS podría facilitar que los resultados obtenidos puedan interpretarse y trasladarse al cuidado de pacientes de una forma adecuada.

Contribuciones de autoría

J. Vidal-Mondéjar diseñó el estudio con ayuda de L. Tejedor-Romero y F. Catalá-López. J. Vidal-Mondéjar y L. Tejedor-Romero obtuvieron los datos. J. Vidal-Mondéjar realizó los análisis. J. Vidal-Mondéjar, L. Tejedor-Romero y F. Catalá-López interpretaron los datos. J. Vidal-Mondéjar

escribió el primer borrador del manuscrito. L. Tejedor-Romero y F. Catalá-López realizaron comentarios, revisiones críticas a las diferentes versiones del texto y editaron la versión final. F. Catalá-López supervisó el estudio. Todos los autores han leído y aprobado la versión final.

Financiación

Los autores declaran que el presente trabajo no ha recibido financiación de forma específica. F. C-L ha recibido ayudas del Instituto de Salud Carlos III/CIBERSAM.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.rx.2023.01.007.

Bibliografía

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. Communicating radiation risks in paediatric imaging: Information to support health care discussions about benefit and risk [consultado 11 Ene 2023]. Disponible en: https://apps.who.int/iris/bitstream/handle/10665/205033/9789241510349_eng.pdf; 2016.
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. Standards for the communication of radiological reports and fail-safe alert notification [consultado 11 Ene 2023]. Disponible en: https://www.rcr.ac.uk/system/files/publication/field_publication_files/bfcr164_failsafe.pdf; 2016.
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. En: Lee G, Fujita H, editores. 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 Di. 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 extensión. 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, Cheng W, 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, and 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;319:388–96, 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, et al. 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.