Inteligencia artificial en el programa de cribado de cáncer de mama

Informes de Evaluación de Tecnologías Sanitarias

INFORMES, ESTUDIOS E INVESTIGACIÓN

MINISTERIO DE SANIDAD

RED ESPAÑOLA DE AGENCIAS DE EVALUACIÓN DE TECNOLOGÍAS Y PRESTACIONES DEL SISTEMA NACIONAL DE SALUD

EUSKO JAURLARITZA
GOBIERNO VASCO
OSASUN SAILA
DEPARTAMENTO DE SALUD

Inteligencia artificial en el programa de cribado de cáncer de mama

Informes de Evaluación de Tecnologías Sanitarias

INFORMES, ESTUDIOS E INVESTIGACIÓN

MINISTERIO DE SANDAD

RED ESPAÑOLA DE AGENCIAS DE EVALUACIÓN DE TECNOLOGÍAS Y PRESTACIONES DEL SISTEMA NACIONAL DE SALUD

EUSKO JAURLARITZA GOBIERNO VASCO
OSASUN SAILA
DEPARTAMENTO DE SALUD

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco

Vitoria-Gasteiz, 2024

Un registro bibliográfico de esta obra puede consultarse en el catálogo de la Biblioteca General del Gobierno Vasco:

https://www.katalogoak.euskadi.eus/katalogobateratua

Edición: 2024

Internet: www.euskadi.eus/publicaciones

Edita: Ministerio de Sanidad
Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco
c/ Donostia-San Sebastián, 1 - 01010 Vitoria-Gasteiz

Fotocomposición: Ipar, S. Coop.
Zurbaran, 2-4 (bajo) — 48007 Bilbao

NIPO: 133-24-098-3

Inteligencia artificial en el programa de cribado de cáncer de mama. Juan Carlos Bayón Yusta, Lorea Galnares Cordero, Iñaki Gutiérrez Ibarluzea. Vitoria-Gasteiz. Ministerio de Sanidad, / Jaurlaritzaren Argitalpen Zerbitzu Nagusia, Servicio Central de Publicaciones del Gobierno Vasco; 2024.

1 archivo pdf; (Informes, Estudios e Investigación)

NIPO: 133-24-098-3

Este documento ha sido realizado por OSTEBA en el marco de la financiación del Ministerio de Sanidad para el desarrollo de las actividades del Plan anual de trabajo de la Red Española de Agencias de Evaluación de Tecnologías y Prestaciones del SNS, aprobado en el Pleno del Consejo Interterritorial del SNS de 26 de mayo de 2021.

Para citar este informe:

Bayón Yusta J C, Galnares Cordero L, Gutiérrez Ibarluzea I. Inteligencia artificial en el programa de cribado de cáncer de mama. Ministerio de Sanidad. Servicio de Evaluación de Tecnologías Sanitarias del País Vasco; 2024. **Informes de Evaluación de Tecnologías Sanitarias:** OSTEBA.

Índice de autores

Grupo de trabajo

Juan Carlos Bayón Yusta. Fundación Vasca de Innovación e Investigación Sanitarias, Gestión del Conocimiento y Evaluación, Bioef-Osteba, Barakaldo, España.

Lorea Galnares Cordero. Fundación Vasca de Innovación e Investigación Sanitarias, Gestión del Conocimiento y Evaluación, Bioef-Osteba, Barakaldo, España.

Iñaki Gutiérrez Ibarluzea. Fundación Vasca de Innovación e Investigación Sanitarias, Coordinador de Gestión del Conocimiento y Evaluación, Bioef-Osteba, Barakaldo, España.

Revisión del Informe

José Luis del Cura Rodríguez. Jefe del Servicio de Radiología. Hospital Universitario de Donostia. Donostia, España.

Cristina Romero Castellano. Jefa del Servicio de Radiodiagnóstico. Complejo Hospitalario Universitario de Toledo. Toledo, España.

María Ederra Sanz. Jefa de Sección Asistencial F.e.a./adjunto. Instituto de la Salud Pública y Laboral de Navarra. Pamplona, España.

Declaración de conflicto de intereses

Los/as autores/as declaran no tener ningún conflicto de intereses en relación con este informe de evaluación.

Desarrollo del proyecto

Desarrollo científico y coordinación técnica: Juan Carlos Bayón Yusta (Bioef-Osteba).

Documentación: Lorea Galnares-Cordero (Bioef-Osteba).

Edición y difusión: Anaitz Leunda Iñurritegui (Bioef-Osteba), Lorea Galnares-Cordero (Bioef-Osteba).

Autor para correspondencia

Juan Carlos Bayón Yusta. jcbayon@bioef.eus
osteba@bioef.eus

Índice

Abreviaturas

ACOG	*American College of Obstetricians and Gynecologist.*
ACR	*American College of Radiology.*
AEPD	Agencia Española de Protección de Datos.
AMSTAR-2	*A Measurement Tool to Assess Systematic Reviews.*
AUC	Área bajo la curva.
AUC-ROC	Área bajo la curva de la curva ROC.
AVAC	Años de vida ganados ajustados por calidad.
AVG	Años de vida ganados.
BI-RADS	Sistema de datos e informes de imágenes mamarias (por sus siglas en inglés).
CAD	Detección asistida por ordenador (por sus siglas en inglés).
CC	Craneocaudal.
CC.AA.	Comunidades autónomas.
CEM	Modelo conjunto de desafío (por sus siglas en inglés).
CNN	Redes neuronales convolucionales (por sus siglas en inglés).
DBT	Mamografía digital con tomosíntesis (por sus siglas en inglés).
DICOM	Imagen digital y comunicaciones en medicina (por sus siglas en inglés).
DL	Aprendizaje profundo (por sus siglas en inglés).
DM	Mamografía digital (por sus siglas en inglés).
DREAM	Iniciativa de diálogo sobre evaluación y métodos de ingeniería inversa (por sus siglas en inglés).
ECA	Ensayos clínicos aleatorizados.
EE	Evaluación económica.
ETS	Evaluación de tecnologías sanitarias.

FDA	Agencia para la Administración de Alimento y Medicamentos en EE. UU. (por sus siglas en inglés).
FFDM	Mamografías digitales de campo completo (por sus siglas en inglés).
FLC	Fichas de lectura crítica.
FN	Falso negativo.
FP	Falso positivo.
IA	Inteligencia artificial.
MA	Metaanálisis.
MQSA	Ley de normas de calidad de la mamografía (por sus siglas en inglés).
ML	Aprendizaje automático (por sus siglas en inglés).
NDS	Nivel de sospecha.
OML	Oblicua mediolateral.
PACS	Sistema de archivo y comunicación de imágenes (por sus siglas en inglés).
PDPCM	Programa de detección precoz de cáncer de mama.
PDM	Probabilidad de malignidad.
PGDPCM	Programa Gallego de Detección Precoz del Cáncer de Mama.
PICO	Población, intervención, comparador, outcome / resultado.
PRS	Puntuación de riesgo poligénico (por sus siglas en inglés).
QUADAS-2	*Quality Assessment of Diagnostic Accuracy Studies.*
RCEI	Ratio coste-efectividad incremental.
RR	Riesgo relativo.
RS	Revisión sistemática.
SERAM	Sociedad Española de Radiología Médica.
SNS	Sistema Nacional de Salud.
UR	Umbral de revaloración.
USPSTF	*United States Preventive Service Task Force.*

Resumen estructurado

Título: Inteligencia artificial en los programas de cribado de cáncer de mama.
Autores: Bayón Yusta JC, Galnares-Cordero L, Gutiérrez Ibarluzea I.
Palabras clave: cáncer de mama, inteligencia artificial, mamografía, cribado, costes.
Fecha: 7/05/2024.
Páginas: 220.
Referencias: 73.
Lenguaje: castellano, y resumen en castellano, euskera e inglés.

Introducción

En España, en la década de los noventa se pusieron en marcha programas para la detección precoz del cáncer de mama (PDPCM), con el objetivo de detectar los cánceres de mama en mujeres en el estadio más precoz posible, y de esta manera, disminuir la mortalidad por dicha causa, mejorar el pronóstico y aumentar la calidad de vida de las mujeres afectadas. A pesar de la buena cobertura y participación observada, surgen dudas sobre la efectividad de los programas, debido al empleo de una tecnología 2D como es la mamografía, de la carga de trabajo añadida de los especialistas en radiología y del sobrediagnóstico y su morbilidad asociada inherente al cribado. Además, no todos los tumores diagnosticados en mujeres sometidas al cribado poblacional se detectan en las revisiones programadas.

En la actualidad hay un interés creciente en el uso de sistemas de inteligencia artificial (IA) en estos cribados. El desarrollo de sistemas de IA como herramientas de apoyo a la detección y diagnóstico radiológico y a la clasificación y notificación radiológica pueden ser una solución a los problemas mencionados de los PDPCM, pudiendo mejorar la detección de lesiones malignas, reducir los cánceres de intervalo, reducir la carga de trabajo de los especialistas en radiología e incluso mejorar la ratio entre coste y beneficio de dichos programas.

Objetivo

Evaluar la eficacia clínica y la eficiencia de incorporar los sistemas de IA a los PDPCM mediante una revisión sistemática de la evidencia científica.

Metodología

Se realizó una búsqueda sistemática de la literatura científica en las siguientes bases de datos: Cochrane Library, International HTA Database, Medline (PubMed), Embase (OvidWed), Web of Science y Scopus, para la identificación de estudios de eficacia y de coste efectividad. Asimismo, se realizó un análisis económico para estimar el coste incremental asociado a la realización del cribado mediante un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble en comparación con la doble lectura estándar. El análisis se realizó desde la perspectiva del financiador del Sistema Nacional de Salud y para un horizonte temporal a corto plazo, y se calculó en base a los costes directos sanitarios y la carga de trabajo de los especialistas en radiología asociados a las estrategias analizadas. Teniendo en cuenta la incertidumbre que pudieran presentar las variables, se realizó un análisis de sensibilidad univariante.

Análisis económico: (**SÍ**) No **Opinión de expertos**: SÍ (**NO**)

Resultados

Se incluyeron una revisión sistemática y once estudios individuales sobre eficacia clínica y un estudio de costes.

Nueve estudios analizaron sistemas de IA como sistemas de lectura autónomo. En cinco la capacidad discriminativa fue buena y en tres, aceptable. Para distintos puntos de corte, en cuatro estudios la sensibilidad de los sistemas de IA fue superior a la de la lectura original única del especialista en radiología.

Ocho estudios analizaron sistemas de IA como ayuda a la lectura radiológica. En uno la capacidad discriminativa fue buena, en cinco aceptable y no aceptable en dos. En los ocho estudios la sensibilidad de la lectura realizada por especialistas en radiología con la ayuda de los sistemas de IA fue superior a la realizada sin ayuda de sistemas de IA.

Siete estudios evaluaron la sensibilidad y especificidad de los sistemas de IA como herramienta de clasificación previa al cribado. En cinco, para umbrales bajos de probabilidad de cáncer visible en la mamografía, la sensibilidad fue alta y en dos, la sensibilidad varió para umbrales altos.

En nueve estudios se analizó el efecto de los sistemas de IA en la carga de trabajo de los especialistas en radiología. En seis (cuatro en los que utilizó como ayuda al especialista en radiología y dos como herramienta de clasificación) se observó una disminución en la carga de trabajo (tiempo de lectura) de los especialistas en radiología.

El estudio de coste-efectividad analizado indicó que la estrategia más coste-efectiva en comparación con el resto de las estrategias analizadas fue la que para el grupo de mujeres entre 40-49 años, utilizó la IA para la predicción inicial del riesgo de cáncer de mama seguida de cribado anual solo para las mujeres de alto riesgo, más cribado de mama a partir de los 50 años hasta los 74, según las directrices USPTF,

Por otro lado, y para una cohorte poblacional de 50.000 mujeres cribadas, se calculó un coste por estudio mamográfico realizado por el sistema de IA y por el especialista en radiología de 0,82 € y 6,39 €, respectivamente, y una reducción en la carga de trabajo (lectura mamográfica de pantalla) del especialista en radiología del 44,3 %, el coste incremental para la estrategia cribado mediante un sistema de IA fue de –253.384,62 €.

Conclusión

La evidencia revisada sugiere que los sistemas de IA son más precisos cuando se emplean en la fase del proceso de cribado como ayuda a la lectura única radiológica y como herramienta de clasificación previa al cribado.

El estudio de evaluación económica analizado señala que realizar a todas las mujeres de 40 años una mamografía índice interpretada mediante IA para predecir el riesgo de cáncer de mama, más cribado anual, desde los 40 a los 49 años, a aquellas en las que se prevé un riesgo elevado de cáncer de mama (RR ≥ 1,1), más cribado de cáncer de mama de acuerdo con las directrices de la United States Preventive Task Force (USPTF), desde los 50 a los 74 años, es una estrategia coste-efectiva.

Para el caso base analizado, el coste asociado a la realización de un cribado poblacional de cáncer de mama mediante una estrategia de cribado que emplea un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura por especialista en radiología única o doble es menor en comparación con la estrategia de doble lectura estándar. El coste ocasionado por el sistema de IA, 41.140 €, queda compensado por el menor coste de lectura de imágenes mamográficas realizada por especialista en radiología en el grupo intervención, consecuencia de realizarse 46.095 lecturas menos en este en comparación con el grupo control.

Finalmente, cabe reseñar que la incorporación de la IA en el cribado de cáncer de mama es un reto con implicaciones éticas, legales y sociales, no solo técnicas, que deben considerarse cuidadosamente para evitar consecuencias perjudiciales para los individuos y grupos, especialmente para los menos favorecidos. De hecho, los sistemas comerciales no ofrecen garantías completas de aplicación a distintas cohortes porque sus desarrollos se han realizado en cohortes y contextos diferentes.

Laburpen egituratua

Izenburua: Adimen artifiziala bularreko minbiziaren baheketa-programetan.
Egileak: Bayón Yusta JC, Galnares-Cordero L, Gutiérrez Ibarluzea I.
Gako-hitzak: bularreko minbizia, adimen artifiziala, mamografia, baheketa, kostuak.
Data: 2024-05-07.
Orrialde kopurua: 220.
Erreferentziak: 73.
Hizkuntza: gaztelania; laburpena gaztelaniaz, euskaraz eta ingelesez.

Sarrera

Espainian, laurogeita hamarreko hamarkadan, bularreko minbizia goiz detektatzeko programak (TMGDP) jarri ziren abian, emakumeengan bularreko minbiziak ahalik eta fase goiztiarrenean detektatzeko eta, horrela, kausa horren ondoriozko hilkortasuna murrizteko, pronostikoa hobetzeko eta kaltetutako emakumeen bizi-kalitatea hobetzeko. Estaldura eta parte-hartze ona izan arren, zalantzak sortzen dira programen eraginkortasunari buruz, 2D teknologia (mamografia) erabiltzen delako, erradiologiako espezialisten lan-karga erantsiagatik eta baheketari lotutako gaindiagnostikoagatik eta erikortasunagatik. Gainera, populazio-baheketa jasan duten emakumeetan diagnostikatutako tumore guztiak ez dira programatutako azterketetan detektatzen.

Gaur egun, gero eta interes handiagoa dago baheketa horietan adimen artifizialeko sistemak (AA) erabiltzeko. Detekzio eta diagnostiko erradiologikorako eta sailkapen eta jakinarazpen erradiologikorako laguntza-tresna gisa AAko sistemak garatzea TMGDPetako aipatutako arazoen konponbidea izan daiteke, lesio gaiztoen detekzioa hobetu, tarteko minbiziak murriztu, erradiologiako espezialisten lan-karga murriztu eta programa horien kostu eta onuren arteko ratioa hobetu baitezake.

Xedea

TMGDPetan AA sistemak txertatzeko eraginkortasun klinikoa eta efizientzia ebaluatzea, ebidentzia zientifikoaren berrikuspen sistematiko baten bidez.

Metodologia

Literatura zientifikoaren bilaketa sistematikoa egin zen datu-base hauetan: Cochrane Library, International HTA Database, Medline (PubMed), Embase (OvidWed), Web of Science eta Scopus, eraginkortasunari eta kostu-eraginkortasunari buruzko azterlanak identifikatzeko. Era berean, analisi ekonomiko bat egin zen baheketa AA sistema baten bidez egiteari lotutako kostu inkrementala zenbatesteko, irakurketa bakarreko edo bikoitzeko baheketa-azterketak irakurketa bikoitz estandarrarekin alderatuta sailkatzeko erabiltzen den detekzioaren lagungarri gisa. Azterketa Osasun Sistema Nazionalaren finantzatzailearen ikuspegitik egin zen, epe laburrerako, eta zuzeneko osasun-kostuen arabera kalkulatu zen, bai eta erradiologiako espezialisten lan-kargaren arabera ere, aztertutako estrategiei lotuta. Aldagaiek izan zezaketen ziurgabetasuna kontuan hartuta, aldagai bakarreko sentikortasun-analisi bat egin zen.

Analisi ekonomikoa: (**BAI**) EZ **Adituen iritzia**: BAI (**EZ**)

Emaitzak

Berrikuspen sistematiko bat eta eraginkortasun klinikoari buruzko hamaika azterlan indibidual sartu ziren, bai eta kostuen azterketa bat ere.

Bederatzi azterlanek AAko sistemak aztertu zituzten irakurketa autonomoko sistema gisa. Bost kasutan, diskriminazio-gaitasuna ona izan zen, eta hirutan, onargarria. Ebaketa-puntu desberdinetarako, lau azterketatan AA sistemen sentikortasuna erradiologiako espezialistaren jatorrizko irakurketa bakarrekoa baino handiagoa izan zen.

Zortzi azterlanek AA sistemak aztertu zituzten irakurketa erradiologikoari laguntzeko. Batean, diskriminazio-gaitasuna ona izan zen, bostetan onargarria eta bitan ez onargarria. Zortzi azterlanetan, erradiologiako espezialistek AAko sistemen laguntzarekin egindako irakurketaren sentsibilitatea AAko sistemen laguntzarik gabe egindakoa baino handiagoa izan zen.

Zazpi azterlanek AA sistemen sentsibilitatea eta espezifikotasuna ebaluatu zuten, baheketa egin aurretik sailkatzeko tresna gisa. Bost kasutan, mamografian ikus zitekeen minbizi-probabilitatearen atalaseetarako, sentikortasuna handia izan zen, eta bitan, sentikortasuna aldatu egin zen atalase altuetan.

Bederatzi azterlanetan, AA sistemek erradiologiako espezialisten lan-kargan duten eragina aztertu zen. Sei kasutan (lautan, erradiologiako

espezialistari laguntzeko erabili zen, eta bitan, sailkapen-tresna gisa.) erradiologiako espezialisten lan-karga (irakurketa-denbora) murriztu egin zela ikusi zen.

Aztertutako kostu-eraginkortasunaren azterketak adierazi zuen aztertutako gainontzeko estrategiekin alderatuz gero, estrategiarik kostu-eraginkorrena honako hau izan zela: 40-49 urte bitarteko emakumeen taldearentzat AAk erabiltzen zuena bularreko arriskuaren hasierako aurreikuspenerako bularreko baheketa eta, ondoren, urteko baheketa arrisku handiko emakumeentzako soilik, eta 50 urtetik 74 urtera arte bularreko baheketa, USPTF jarraibideen arabera.

Bestalde, eta 50.000 emakume bahetuko populazio-kohorte baterako, AA sistemak eta erradiologiako espezialistak egindako azterketa mamografiko bakoitzeko 0,82 € eta 6,39 €-ko kostua kalkulatu zen, hurrenez hurren, eta erradiologiako espezialistaren lan-karga % 44,3 murriztu zen (pantaila-irakurketa mamografikoa); AA sistema baten bidezko baheketa-estrategiarako kostu inkrementala –253.384,62 € izan zen.

Ondorioa

Berrikusitako ebidentziak iradokitzen du AAko sistemak zehatzagoak direla baheketa-prozesuaren fasean irakurketa erradiologiko bakarrari laguntzeko eta baheketa egin aurretik sailkatzeko erabiltzen direnean.

Ebaluazio ekonomikoaren azterketaren arabera, kostu-eraginkorreko estrategia da 40 urteko emakume guztiei AA bidez interpretatutako mamografia indizea egitea, bularreko minbiziaren arriskua aurreikusteko. Gainera, urtean behin baheketa bat egin behar zaie 40-49 urte bitarteko emakumeei, bularreko minbizia izateko arrisku handia aurreikusten bada (RR ≥ 1,1). Era berean, bularreko minbiziaren baheketa egin behar da 50 eta 74 urte bitartean, United States Preventive Task Force (USPTF) erakundearen jarraibideen arabera.

Aztertutako oinarrizko kasurako, AAko sistema bat erabiltzen duen baheketa-estrategia baten bidez bularreko minbiziaren populazio-baheketa bat egiteari lotutako kostua —baheketa-azterketak sailkatzeko eta erradiologia bakarreko edo bikoitzeko espezialistak irakurtzeko erabiltzen dena— txikiagoa da irakurketa estandar bikoitzeko estrategiarekin alderatuta. AA sistemak eragindako kostua, 41.140 €, konpentsatu egiten da parte-hartze taldean erradiologiako espezialistak egindako mamografia-irudien irakurketaren kostu txikiagoarekin, 46.095 irakurketa gutxiago egiten baitira kontrol-taldearekin alderatuta.

Azkenik, adierazi behar da bularreko minbiziaren baheketan AA sartzea erronka bat dela, inplikazio etiko, legal eta sozialak dituena, ez soilik teknikoak, eta kontu handiz aztertu behar direla, gizabanakoentzat eta taldeentzat, batez ere behartsuenentzat, ondorio kaltegarriak saihesteko. Izan ere, merkataritza-sistemek ez dute hainbat kohorteri aplikatzeko berme osorik eskaintzen, euren garapenak kohorte eta testuinguru desberdinetan egin direlako.

Structured summary

Title: The use of artificial intelligence in breast cancer screening programmes.
Authors: Bayón Yusta JC, Galnares-Cordero L, Gutiérrez Ibarluzea I.
Keywords: breast cancer, artificial intelligence, mammography, screening, costs.
Date: 7 May 2024.
Number of pages: 220.
References: 73.
Languages: Spanish and abstract in Spanish, Basque and English.

Introduction

Breast Cancer Early Detection Programmes (BCEDPs) were set up in Spain in the 1990s, aiming to diagnose breast cancer in women as early as possible and thereby reduce the associated mortality, improve prognosis and increase patient quality of life. Despite good coverage and participation rates, there are concerns about the effectiveness of such programmes, due to the use of 2D technology as employed in mammography, the extra workload for radiologists, overdiagnosis and screening-related morbidity. Further, not all the tumours diagnosed in women who undergo population screening are detected in the scheduled check-ups.

There is a growing interest in using artificial intelligence (AI) systems for this type of screening. The development of AI systems for assisting radiological detection, diagnosis, classification and notification may help tackle the problems associated with BCEDPs, potentially improving the detection of malignancy, reducing interval cancer rates, lessening the workload for radiologists and even improving the cost benefit ratio of screening programmes.

Objective

To assess the clinical efficacy and efficiency of adopting AI systems in BCEDPs through a systematic review of the scientific evidence.

Methodology

To identify studies on efficacy and cost-effectiveness, a systematic review of the scientific literature was performed using the following databases: Cochrane Library, International HTA Database, Medline

(PubMed), Embase (Ovid Web), Web of Science and Scopus. In addition, economic analysis was conducted to estimate the incremental costs associated with screening with an AI support tool for detection, using this tool to assign screening exams to single or double reading compared to standard double reading. The analysis was carried out from the perspective of the funding body of the Spanish National Health System and with a short time horizon, and costs were estimated using direct healthcare costs and the workload of radiologists associated with the strategies analysed. Given potential uncertainties in the data, univariate sensitivity analysis was also performed.

Economic analysis: (**YES**) NO **Expert opinion**: YES (**NO**)

Results

Overall, 1 systematic review and 11 individual studies on clinical efficacy and 1 cost-analysis study were included.

Nine studies analysed AI-based automated reading systems. The discriminatory power of these systems was considered to be good in five of the studies and acceptable in three. Four studies reported that the AI systems were more sensitive than a single reading by a radiologist for certain cut-off points.

Eight studies analysed AI systems as tools to support reading of radiological images. The discriminatory power was considered good in one case, acceptable in five and unacceptable in two. In all eight studies, readings by radiologists were more sensitive when made with than without AI assistance.

Seven studies assessed the sensitivity and specificity of AI systems as a classification tool before screening. Sensitivity was high for lower thresholds for the risk of cancer detection using mammography in five studies, while it was variable for higher risk thresholds in two studies.

Nine studies analysed the impact of AI systems on specialists' workload. Six studies (four considering AI as a support tool for the radiologist and two as a classification tool) reported a reduction in radiologists' workload (reading time).

The cost-effectiveness study included indicated that the most cost-effective strategy among those analysed was the one that used AI for the initial prediction of breast cancer for the 40- to 49-year-old age group, followed by annual screening only for the women in this group classified as high risk, and then screening from 50 up to 74 years of age, in line with the United States Preventive Services Task Force (USPSTF) guidelines.

On the other hand, for a population cohort of 50,000 screened women, it was calculated that the costs per mammography exam carried out using an AI-based system and by a radiologist were €0.82 and €6.39, respectively, with a 44.3 % reduction in radiologists' workload (mammography screen reading), and overall, the incremental cost for a screening strategy using an AI system was –€253,384.62.

Conclusion

The evidence reviewed suggests that AI systems are most accurate when used in the screening process as a tool to support single reading of mammograms and pre-screening triage.

The economic analysis indicates that the following is a cost-effective strategy: obtaining an index mammography in all women aged 40 years old that is then interpreted by AI to predict breast cancer risk, plus annual screening from 40 to 49 years of age in women predicted to be at high risk of breast cancer (relative risk ≥ 1.1), followed by breast cancer screening from 50 to 74 years old in accordance with the USPSTF guidelines.

For the base case scenario considered, the costs of breast cancer population screening with a strategy based on an AI system to assist detection, using it to classify exams before single or double reading by specialists, are lower than the standard double reading procedure. The costs associated with the AI system (€41,140) were compensated for by the lower costs associated with mammography reading by radiologists in the intervention group (which required 46,095 fewer readings than the control group).

Finally, we should highlight that the use of AI in breast cancer screening is a challenge with ethical, legal and social as well as technical implications, and that should be considered carefully to avoid harmful effects for individuals and groups, especially the most disadvantaged. Indeed, as commercial systems have been developed based on research in specific cohorts and contexts, there is no guarantee that they are applicable to other populations.

I. Introducción y justificación

El cáncer más frecuentemente diagnosticado a nivel mundial en el año 2020 fue el de mama, con una incidencia del 12,5 % (2.261.419 casos). En España, se estima que el cáncer de mama será el segundo más frecuentemente diagnosticado en el año 2024, 36.395 nuevos casos, el primero en mujeres. Además, en el año 2020 a nivel mundial se estimó una prevalencia a los cinco años del diagnóstico del cáncer de mama del 17,7 % (7.790.717 casos). En España, para el mismo año, se estimó en mujeres una prevalencia total de 516.827 casos y una prevalencia a los cinco años de 144.233 casos. Por otro lado, en el año 2020 a nivel mundial el cáncer de mama fue responsable de 684.996 fallecimientos (6,9 % del total de fallecimientos por cáncer). En España, en el año 2022, el número de fallecimientos totales por tumor maligno de la mama fue de 6.754, siendo el tumor que más muertes causó en mujeres, 6.677 casos. Por último, la supervivencia neta a cinco años del diagnóstico de las pacientes diagnosticas en el periodo 2008-2013 de cáncer mama fue del 85,5 % (1).

Dada la elevada carga del cáncer de mama, en España, en la década de los noventa, las comunidades autónomas (CC.AA.) y ciudades autónomas pusieron en marcha programas para la detección precoz del cáncer de mama (PDPCM), con el objetivo de detectar los cánceres de mama en el estadio más precoz posible, con el fin de disminuir la mortalidad por dicha causa, mejorar el pronóstico y aumentar la calidad de vida de las mujeres afectadas. Estos programas poblacionales de cribado de mama, dirigidos a mujeres de un grupo de edad de entre 50-69 años o de entre 45-69 años, en 2017 alcanzaron una cobertura del 88,98 % de la población objetivo y una tasa de participación del 75,66 % (2).

A pesar de la buena cobertura y participación observada, surgen dudas sobre la efectividad de los PDPCM debido al empleo de una técnica 2D como es la mamografía, que genera solapamiento de tejidos y reduce la capacidad de detección, y a la alta carga de trabajo de los especialistas en radiología, derivada del elevado número de mamografías y del procedimiento de doble lectura adoptado para aumentar la tasa de detección de cáncer y que últimamente se ve obstaculizado por la falta creciente de especialistas en radiología (3). Otro aspecto controvertido de los PDPCM es el sobrediagnóstico aparejado al cribado, es decir, los cánceres correctamente diagnosticados, pero que por su evolución natural no hubieran causado problemas médicos antes de la muerte de la mujer por otras causas. En estudios realizados por el Cancer Intervention and Surveillance

Modeling Network (CISNET) del National Cancer Institute se estima que una de cada ocho mujeres entre 50 y 75 años serán sobrediagnosticadas, cifra que aumenta si el cribado comienza en edades más tempranas y en mujeres sin antecedentes personales o familiares previos (4). Además, hay que considerar que no todos los tumores diagnosticados en mujeres sometidas al cribado poblacional se detectan en las revisiones programadas. Los programas de cribado mamario pasan por alto entre el 15 y el 35 % de los cánceres, ya sea por error o porque el cáncer no es visible o perceptible para el especialista en radiología. Algunos de estos cánceres no detectados se presentan sintomáticamente como cánceres de intervalo (5).

En los 90 aparecieron los primeros sistemas de detección asistida por ordenador (CAD, por sus siglas en inglés) como una herramienta de apoyo al diagnóstico en mamografía, con el objetivo de procesar automáticamente las imágenes y señalar las áreas consideradas sospechosas. Se han desarrollado dos categorías de algoritmos de CAD, los utilizados para detectar la presencia de una lesión y los que, además, indican si la lesión es benigna o maligna. El mayor inconveniente de los CAD es su moderada especificidad; marca muchas áreas como sospechosas, lo que conduce a falsos positivos (FP) o falsos negativos (FN).

Los recientes avances en inteligencia artificial (IA) proporcionan una serie de posibilidades que van más allá de las ofrecidas por el CAD tradicional en el apoyo a los especialistas en radiología durante la lectura de los exámenes de mamografía en los PDPCM. Los objetivos de los sistemas de IA son los mismos que los del CAD tradicional: mejorar la detección de lesiones malignas, reducir cánceres de intervalo y, al mismo tiempo, reducir la carga de lectura. Eventualmente, es posible que los nuevos sistemas y técnicas de la IA puedan incluso mejorar la ratio entre coste y beneficio de los PDPCM (3).

Se afirma que el reconocimiento de imágenes mediante IA para el cribado mamario podría mejorar el de los especialistas en radiología experimentados y solventar algunas de las limitaciones señaladas de los PDPCM. Por ejemplo, podrían perderse menos cánceres porque un algoritmo de IA no se ve afectado por la fatiga y el diagnóstico subjetivo de los especialistas en radiología y podría disminuir la carga de trabajo o sustituir por completo a los especialistas en radiología. Sin embargo, también podría aumentar los efectos indeseados del cribado. Así, la IA podría alterar el espectro de la enfermedad detectada en el cribado mamario si detecta de forma diferencial más microcalcificaciones, que se asocian a un carcinoma ductal *in situ* de menor grado, lo que incrementaría las tasas de sobrediagnóstico y sobretratamiento y alteraría el balance entre beneficio y daño (5).

Un sistema de IA con una especificidad mayor que la del especialista en radiología tendría un efecto importante sobre el número de revaloraciones por FP en un PDPCM, ya que cada una de estas causa un daño a la mujer cribada y supone un coste financiero y de mano de obra al Sistema Nacional de Salud (SNS). Del mismo modo, un sistema de IA capaz de detectar el mismo tipo de cáncer o un tipo de cáncer más agresivo con una sensibilidad similar a la del especialista en radiología sería preferible a un sistema de IA con una sensibilidad superior que detectase pacientes adicionales que, sin embargo, fuesen predominantemente diagnosticadas con carcinoma ductal *in situ* con bajo grado. Adicionalmente, si un sistema de IA fuera sustancialmente más sensible que la práctica actual, sería importante disponer de pruebas sobre si esa detección adicional de pacientes con cáncer en el cribado conduce a un menor número de cánceres sintomáticos, a un menor número de cánceres en estadios avanzados (6) o a una mayor detección de cánceres potencialmente mortales, lo que podría compensar el sobrediagnóstico. Es decir, en una enfermedad potencialmente mortal como el cáncer de mama, la pregunta sería saber cuántas vidas salvadas vale el coste añadido del sobrediagnóstico.

Además, el espectro de la enfermedad también podría estar condicionado por los conjuntos de casos utilizados en el entrenamiento de los sistemas de IA y por las estructuras del sistema. Las estructuras y algoritmos de los sistemas de IA no son siempre transparentes o explicables por lo que su interpretación puede ser un problema. Cómo y porqué un algoritmo toma una decisión puede ser difícil de entender ya que los algoritmos no comprenden el contexto, el modo de recogida o el significado de las imágenes vistas, lo que puede provocar que las redes neuronales profundas lleguen a la conclusión de un problema a través de un atajo, en lugar de la solución prevista. Esto refleja la importancia de evitar posibles factores de confusión en los datos de entrenamiento, la comprensión de la toma de decisiones mediante algoritmos y el papel fundamental de una evaluación rigurosa (5).

I.1. Detección asistida por ordenador tradicional

Los sistemas tradicionales de CAD buscan en la imagen características específicas de las lesiones previamente definidas por los especialistas en radiología (3). Estos sistemas, están entrenados para detectar microcalcificaciones y masas, mediante la observación de una forma, textura, gradiente y nivel de gris específicos (7).

Los sistemas CAD están divididos en sistemas de detección y sistemas de diagnóstico asistidos por ordenador. Los sistemas de detección identifican anomalías sospechosas, pero dejan en manos de los especialistas en radiología la decisión sobre la gestión de la paciente. Los sistemas de diagnóstico estiman la probabilidad de la enfermedad con base en las características de la anomalía (7). El mayor inconveniente de los sistemas tradicionales de CAD es su moderada especificidad y el mayor tiempo empleado en la lectura de las imágenes, dado que estos marcan muchas áreas como sospechosas que conducen a FP o FN por parte de los especialistas en radiología (3) y a revaloraciones innecesarias para las pacientes (7). Además, los sistemas de CAD tradicionales tienen una mayor sensibilidad para la detección de microcalcificaciones que para la detección de masas, distorsiones arquitecturales y asimetrías, que están englobadas en la misma marca a pesar de ser hallazgos diferentes. Esto hace que cuando se manifiestan como microcalcificaciones la sensibilidad del CAD sea mayor del 95 %, siendo del 80 % cuando lo hace como masa y del 50 % en distorsiones o asimetrías (8).

I.2. Detección asistida por ordenador basada en redes neuronales convolucionales

La IA estudia la incorporación de comportamientos de la inteligencia humana a las máquinas, por ejemplo, el aprendizaje. El elemento que provee a la IA la capacidad de aprender una tarea a partir de experiencias previas sin necesidad de una programación específica es el aprendizaje automático (ML, por sus siglas en inglés). Un subconjunto del ML es el aprendizaje profundo (DL, por sus siglas en inglés), el cual comprende técnicas de aprendizaje automático en las que el algoritmo aprende por sí solo las características más importantes de una imagen para la realización de una tarea predeterminada, lo que mejora el rendimiento de la CAD tradicional (3).

Las arquitecturas de aprendizaje más utilizadas actualmente en DL son las redes neuronales, que corresponden a nodos interconectados formando múltiples capas que imitan redes neuronales biológicas cerebrales, y más concretamente las redes neuronales convolucionales (CNN, por sus siglas en inglés) que son un subconjunto de algoritmos basados en operaciones de convolución (3). Las CNN utilizan redes neuronales para aprender de los datos y crear distintas representaciones de ellos (7) sin necesidad de un conocimiento previo ni, por tanto, de la intervención humana, lo

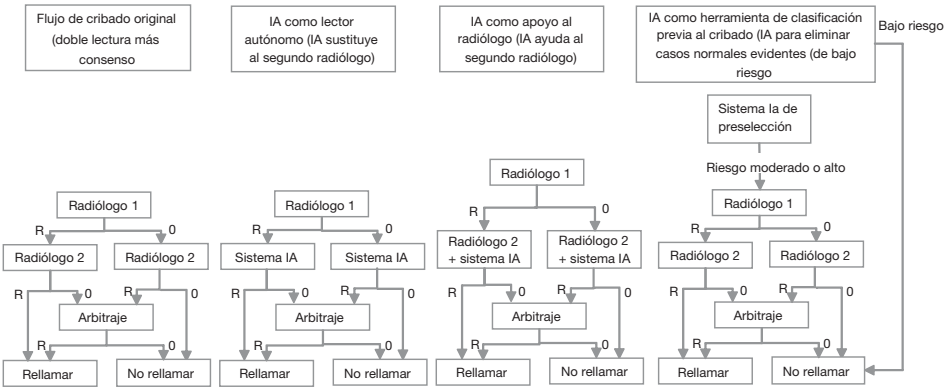
que constituye la diferencia fundamental entre las CNN y los CAD tradicionales (3).

El CAD basado en CNN en comparación con CAD tradicional ha mostrado una tasa menor de FP, lo que implica una reducción de las tasas de revaloración y por tanto de la ansiedad y angustia asociada a esta (7).

I.3. Papel potencial de la inteligencia artificial en el cribado de cáncer de mama

Los sistemas de IA pueden jugar un papel potencial en distintas fases del proceso de cribado de mama. Taylor-Phillips *et al.* (6) propone una serie de escenarios en los que describen cómo la IA puede actuar en estas fases. Cómo se propone utilizar los sistemas de IA y cómo interactúan los especialistas en radiología con estos sistemas son aspectos a tener en cuenta en la evaluación. La combinación de expertos e IA en la práctica clínica es lo que determinará la precisión global y los resultados para las mujeres (6).

Principalmente, los sistemas de IA pueden ser utilizados en el cribado de mama como instrumento de apoyo para los especialistas en radiología, como lector autónomo o como herramienta de clasificación previa. En la figura 1, tomada del estudio de Taylor-Phillips *et al.* (6) quedan reflejados los roles potenciales de la IA en el proceso de cribado de mama.



0 = recomendación de no rellamar para más pruebas. R = recomendación de rellamar para más pruebas porque hay indicaciones de cáncer.

Figura 1. **Papeles potenciales de la IA en el proceso de cribado de mama**

La IA como instrumento de apoyo a los especialistas en radiología puede tener una influencia directa en su comportamiento proporcionando indicaciones en las imágenes mamográficas con el objetivo de ayudarles (6). Puede asistir al especialista en radiología en la interpretación y proporcionar otra evaluación de la misma interpretación mamográfica. La nueva generación de sistemas CAD basado en IA además de mostrar marcas en regiones sospechosas de las imágenes mamográficas, también proporciona un soporte interactivo o una puntuación a todo el examen indicativa de la sospecha de cáncer de mama (3). La IA como ayuda del especialista en radiología puede utilizarse para apoyar la decisión del primer especialista en radiología, del segundo especialista en radiología o el arbitraje, o ambos (6).

La IA como lector autónomo puede sustituir a los especialistas en radiología al tener capacidad de categorizar las mamografías según su probabilidad de contener cáncer. Los casos clasificados por el sistema de IA como sospechosos son revisados por los especialistas en radiología en primer lugar, evitando así que las lesiones no sean detectadas y favoreciendo que la mujer sea revalorada lo antes posible. Teniendo en cuenta esta capacidad y que más del 95 % de las mujeres cribadas no tienen ningún tipo de anormalidad, es posible que se pueda utilizar la IA como primer lector. El papel que jugaría sería clasificar automáticamente los exámenes de mamografía, lo que podría reducir la carga de trabajo de los especialistas en radiología al evitar que los exámenes normales tuviesen que ser leídos por dos especialistas en radiología, si se realiza doble lectura. En este contexto, el reto mayor para un sistema de IA sería proporcionar automáticamente una decisión de revaloración (6).

Como herramienta de clasificación previa al cribado, la IA se aplica para eliminar todos los casos normales evidentes, es decir los de bajo riesgo. El sistema de IA lee todos los exámenes mamográficos y proporciona una puntuación de examen, generalmente de 1 a 10, que indica la probabilidad creciente de que exista un cáncer visible en la mamografía. Con base en umbrales predeterminados, los exámenes mamográficos clasificados con riesgo moderado o sospechosos de cáncer se remitirían a lectura radiológica, primer lector, segundo lector o consenso, para tomar la decisión de revalorar o no a las mujeres, mientras que los clasificados como normales no supondrían una revaloración (6).

I.4. Sistemas de inteligencia artificial para el cribado de mama aprobados por la FDA

Con fecha de 31 de marzo de 2023, la agencia para la Administración de Alimentos y Medicamentos de EE. UU. (FDA, por sus siglas en inglés) tiene aprobados seis sistemas de IA para la detección y diagnóstico radiológico asistido por ordenador de lesiones en mamografías de cribado: MammoScreen 2.0 (9), Genius AI Detection 2.0 (10), ProFound AI Software V3.0 (11), Transpara 1.7.0 (12), Lunit INSIGHT MMG (13) y Saige-Dx (14); y tres sistemas de IA para la clasificación y notificación radiológica asistida por ordenador de lesiones en mamografías de cribado: cmTriage (15), HealthMammo (16) y ConNet Qm Triage (17) (ver tablas 1 y 2).

Los sistemas de IA para la detección y diagnóstico asistido por ordenador están pensados para ser utilizados como ayuda de lectura simultánea para asistir a los especialistas en radiología con la caracterización, localización y detección de posibles anomalías sospechosas de cáncer de mama y para mejorar su flujo de trabajo. Los algoritmos de IA proporcionan marcas para resaltar el lugar en los que el dispositivo detecta calcificaciones sospechosas o lesiones de tejido blando y puntuaciones de nivel de sospecha correlativa, generalmente de 1 a 100, además de puntuaciones a nivel de mama y/o examen. Los usuarios a los que van destinados los sistemas son médicos cualificados que interpretan mamografías de cribado, bien mamografías digitales de campo completo (FFDM, por sus siglas en inglés) o bien mamografías digitales con tomosíntesis (DBT, por sus siglas en inglés). Dos de los seis sistemas (MammoScreen 2.0 y Transpara 1.7.0) se pueden utilizar tanto para FFDM o DBT, tres (AI Detection 2.0, ProFound AI Software V3.0 y Saige-Dx) para DBT y el algoritmo Lunit INSIGHT MMG para FFDM. Por último, hay que señalar que la población diana en la que se utilizan dichos dispositivos son mujeres sometidas a mamografías de cribado.

Los sistemas de IA para la clasificación y notificación radiológica asistida por ordenar son una herramienta de software de flujo de trabajo paralelo de notificación pasiva que utilizan los especialistas en radiología para priorizar a las pacientes con hallazgos sospechosos de cáncer de mama en un estudio. Los tres sistemas (cmTriage, HealthMammo y ConNet Qm Triage) utilizan un algoritmo de IA para analizar mamografías de cribado de FFDM que marca aquellas imágenes mamográficas que sugieren la presencia al menos de un hallazgo sospechoso a nivel de examen. Los sistemas se limitan a la categorización de los exámenes y no proporcionan ninguna información diagnóstica ni eliminan imágenes de la lista de trabajo del médico intérprete, que es el usuario de éstos.

Tabla 1. **Dispositivos de IA de detección y diagnóstico aprobados por la FDA en mamografías de cribado**

Dispositivo (Compañía)	MammoScreen® 2.0 (Therapixel) (9)	Genius AI Detection 2.0 (Hologic) (10)	ProFound AI® Software V3.0 (ICAD) (11)
Número / nombre de regulación	21 CFR 892.2090. Software de detección y diagnóstico radiológicos asistidos por ordenador.	21 CFR 892.2090 Software de detección y diagnóstico radiológicos asistidos por ordenador.	21 CFR 892.2090 Software de detección y diagnóstico radiológicos asistidos por ordenador.
Descripción del dispositivo	MammoScreen 2.0 procesa automáticamente las cuatro vistas (una cráneo caudal (CC) y una oblicua mediolateral (OML) por mama) del cribado estándar en FFDM o DBT. Emite un informe en el que proporciona un nivel general de sospecha de cada mama y de la mamografía (puntuación de 1 a10, 10 máxima sospecha) e indicaciones visuales explícitas cuando se detectan exámenes altamente sospechosos.	Genius AI Detection identifica posibles anomalías en las imágenes DBT. Para cada lesión detectada produce resultados que incluyen la ubicación de la lesión, un contorno de la lesión y una puntuación de confianza para esa lesión. También produce una puntuación de caso para todo el examen de tomosíntesis.	ProFound AI® V3.0 detecta densidades de tejido blando maligno y calcificaciones en imágenes DBT. Marca las áreas detectadas en las imágenes DBT, asignando a cada hallazgo y a cada caso una puntuación de 0 a 100 (100 máxima sospecha) para indicar la confianza en que el hallazgo es maligno y de que el caso tiene hallazgos malignos.

Dispositivo (Compañía)	MammoScreen® 2.0 (Therapixel) (9)	Genius AI Detection 2.0 (Hologic) (10)	ProFound AI® Software V3.0 (ICAD) (11)
Indicaciones de uso	MammoScreen® está diseñado para su uso como ayuda de lectura simultánea para los médicos intérpretes, para ayudarles a identificar los hallazgos, que pueden ser tejidos blandos o calcificaciones en FFDM o DBT adquiridas con sistemas compatibles y evaluar su nivel de sospecha. Las decisiones de tratamiento de la paciente no deben tomarse únicamente sobre la base del análisis de MammoScreen.	Genius AI Detection pretende ayudar en la interpretación de los exámenes DBT de mama de forma concurrente, donde el médico intérprete confirma o descarta los hallazgos durante la lectura del examen. Identifica y marca regiones de interés, densidades de tejido blando y calcificaciones en exámenes de DBT de sistemas DBT compatibles, y evalúa la certeza de los hallazgos.	ProFound AI® V3.0 se utiliza simultáneamente por médicos intérpretes mientras leen exámenes de DBT de sistemas DBT compatibles. Detecta densidades de tejido blando y calcificaciones y proporciona puntuaciones de certeza de hallazgo y de caso, lo que ayuda a los médicos intérpretes a identificar las densidades de tejido blando y calcificaciones para ser confirmadas o descartadas por el médico intérprete.
Usuarios	Especialistas en radiología que interpretan mamografía de cribado FFDM o DBT.	Especialistas en radiología cualificados por la ley de normas de calidad de las mamografías (MQSA por sus siglas en inglés).	Especialistas en radiología.
Población diana	Mujeres sometidas a mamografía de cribado FFDM o DBT.	Mujeres sintomáticas y asintomáticas sometidas a mamografía de cribado FFDM yDBT.	Mujeres sintomáticas y asintomáticas sometidas a mamografía.
Tipo de mamografía	DM y DBT	DM y DBT	DBT
Proveedor mamografía	GE Healthcare (DM), Hologic (DM y DBT).	Hologic.	GE Healthcare, Hologic, Siemens Healthineers.
Análisis clínicos: diseño del estudio de lectores	Diseño cruzado multilector-multicaso con una muestra enriquecida.	Estudio multilector-multicaso.	Diseño cruzado multilector-multicaso.

Dispositivo (Compañía)	MammoScreen® 2.0 (Therapixel) (9)	Genius AI Detection 2.0 (Hologic) (10)	ProFound AI® Software V3.0 (ICAD) (11)
Análisis clínico: número de casos analizados	240 casos para DM y 240 casos para DBT para el estudio de la IA como apoyo al especialista en radiología. 240 casos para DM y 240 casos para DBT para el estudio de la IA como lector autónomo.	390 casos para el estudio de la IA como apoyo al especialista en radiología.	240 casos para el estudio de la IA como apoyo al especialista en radiología.
Análisis clínico: número de especialistas en radiología en el estudio del lector	14 para el estudio 2D (DM). 20 para el estudio 3D (DBT).	17 para el estudio 3D (DBT).	24 para el estudio 3D (DBT).
Análisis clínico: resultados	**MammoScreen 2.0 como apoyo a los especialistas en radiología en FFDM:** *Sin apoyo de la IA*: AUC = 0,77. *Con apoyo de la IA*: AUC = 0,80. **MammoScreen 2.0 como apoyo a los especialistas en radiología en DBT:** *Sin apoyo de la IA*: AUC = 0,79. *Con apoyo de la IA*: AUC = 0,83. **MammoScreen 2.0 como lector autónomo en FFDM:** AUC = 0,79. **MammoScreen 2.0 como lector autónomo en DBT:** AUC = 0,84.	**Genius AI Detection 2.0 como apoyo a los especialistas en radiología en DBT:** *Sin apoyo de la IA*: AUC = 0,794. *Con apoyo de la IA*: AUC = 0,825.	**ProFound AI V3.0 como apoyo a los especialistas en radiología en DBT:** *Sin apoyo de la IA*: AUC = 0,795. *Con apoyo de la IA*: AUC = 0,852.

Dispositivo (Compañía)	Transpara 1.7.0 (ScreenPoint Medical) (12)	Lunit INSIGHT MMG (Lunit) (13)	Saige-Dx (DeepHealth) (14)
Número / nombre de regulación	21 CFR 892.2090 Software de detección y diagnóstico radiológicos asistido por ordenador.	21 CFR 892.2090 Software de detección y diagnóstico radiológicos asistidos por ordenador.	21 CFR 892.2090 Software de detección y diagnóstico radiológicos asistidos por ordenador.
Descripción del dispositivo	Transpara® 1.7.0 está diseñado para que los médicos mejoren la interpretación de la FFDM y la DBT. Proporciona: marcas de los lugares en los que detecta calcificaciones sospechosas o lesiones de tejidos blandos, puntuaciones del nivel de sospecha (de 0 a 100, 100 máxima sospecha), enlaces entre regiones correspondientes en diferentes vistas de la mama y una puntuación de examen de 1 al 10 (10 máxima sospecha).	Lunit INSIGHT MMG es un software radiológico para ayudar a los médicos intérpretes en la detección, localización y caracterización de áreas sospechosas de cáncer de mama en mamografías de sistemas FFDM compatibles. Analiza las imágenes recibidas e identifica y caracteriza áreas sospechosas de cáncer de mama. Presenta una puntuación (de 1 a 100), que refleja la probabilidad general de presencia de malignidad, para cada lesión y mama.	Saige-Dx es un dispositivo que ayuda a mejorar el rendimiento de los lectores y reducir el tiempo de lectura. Detecta la presencia o ausencia de lesiones de tejidos blandos y calcificaciones en DBT. Produce cuadros delimitadores que circunscriben cualquier hallazgo detectado y asigna un nivel de sospecha de que el hallazgo sea maligno y un nivel de sospecha de malignidad en todo el caso.

Dispositivo (Compañía)	Transpara 1.7.0 (ScreenPoint Medical) (12)	Lunit INSIGHT MMG (Lunit) (13)	Saige-Dx (DeepHealth) (14)
Indicaciones de uso	El software Transpara® está pensado para su uso como ayuda de lectura simultánea para médicos que interpretan exámenes de FFDM y DBT de sistemas compatibles, para identificar regiones sospechosas de cáncer de mama y evaluar su probabilidad de malignidad. Las decisiones de tratamiento de las pacientes no deben tomarse únicamente sobre la base del análisis de Transpara®.	Lunit INSIGHT MMG está destinado a ayudar en la detección, localización y caracterización de áreas sospechosas de cáncer de mama en mamografías de sistemas FFDM compatibles y como herramienta complementaria, a ser visualizado por los médicos intérpretes después de completar su lectura inicial. No sustituye la revisión completa de un médico o su juicio clínico.	Saige-Dx está pensado para su uso como ayuda de lectura simultánea para los médicos intérpretes en mamografías de cribado con hardware DBT compatible. Analiza mamografías DBT para identificar la presencia o ausencia de lesiones de tejidos blandos y calcificaciones que puedan ser indicativas de cáncer.
Usuario	Médicos cualificados que interpretan mamografía de cribado FFDM o DBT.	Médicos cualificados que interpretan mamografía de cribado FFDM.	Médicos cualificados que interpretan mamografía de cribado DBT.
Población diana	Mujeres sometidas a mamografía de cribado FFDM o DBT.	Mujeres sometidas a mamografía de cribado FFDM y DBT.	Mujeres ≥ 35 años que se sometan a mamografía de cribado DBT.
Tipo de mamografía	DM y DBT	DM y DBT	DBT
Proveedor mamografía	GE Healthcare (DM), Philips Healthcare (DM), Fujifilm (DM y DBT), Hologic (DM y DBT), Siemens Healthineers (DM y DBT).	GE Healthcare, Hologic, Siemens Healthineers.	GE Healthcare, Philips Healthcare, Fujifilm, Hologic, Siemens Healthineers.
Análisis clínicos: diseño del estudio de lectores	Estudio con muestra enriquecida.	Estudio multilector-multicaso.	Estudio multilector-multicaso totalmente equilibrado.

Dispositivo (Compañía)	Transpara 1.7.0 (ScreenPoint Medical) (12)	Lunit INSIGHT MMG (Lunit) (13)	Saige-Dx (DeepHealth) (14)
Análisis clínico: número de casos analizados	240 casos para DM y 240 casos para DBT para el estudio de la IA como apoyo al especialista en radiología. 9.122 casos para el estudio de la IA como lector autónomo.	240 casos para el estudio de la IA como apoyo al especialista en radiología. 240 casos para el estudio de la IA como lector autónomo.	240 casos para el estudio de la IA como apoyo al especialista en radiología. 1.304 casos para el estudio de la IA como lector autónomo
Análisis clínico: número de especialistas en radiología en el estudio del lector	14 para el estudio 2D (DM). 18 para el estudio 3D (DBT).	12 para el estudio 2D (DM).	18 para el estudio 3D (DBT).
Análisis clínico: resultados	**Transpara 1.7.0 como apoyo a los especialistas en radiología en FFDM:** *Sin apoyo de la IA*: AUC = 0,87. *Con apoyo de la IA*: AUC = 0,89. **Transpara 1.7.0 como apoyo a los especialistas en radiología en DBT:** *Sin apoyo de la IA*: AUC = 0,83. *Con apoyo de la IA*: AUC = 0,86. **Transpara 1.7.0 como lector autónomo en FFDM:** AUC = 0,949. **Transpara 1.7.0 como lector autónomo en DBT:** AUC = 0,931.	**Lunit INSIGHT MMG como apoyo a los especialistas en radiología en FFDM:** *Sin apoyo de la IA*: AUC = 0,754. *Con apoyo de la IA*: AUC = 0,805. **Lunit INSIGHT MMG como lector autónomo en DBT:** AUC = 0,863.	**Saige-Dx como apoyo a los especialistas en radiología en DBT:** *Sin apoyo de la IA*: AUC = 0,865. *Con apoyo de la IA*: AUC = 0,925. **Saige-Dx como lector autónomo en DBT:** AUC = 0,930.

AUC = área bajo la curva ROC.

Tabla 2. **Dispositivos de IA de clasificación aprobados por la FDA para mamografías de cribado**

Dispositivo (Compañía)	cmTriage (CureMetrix) (15)	HealthMammo (Zebra Medical Vision) (16)	CogNet QmTRIAGE (Med Cognetics) (17)
Número / nombre de regulación	21 CFR 892.2080 Software de clasificación y notificación radiológica asistida por ordenador.	21 CFR 892.2080 Software de clasificación y notificación radiológica asistida por ordenador.	21 CFR 892.2080 Software de clasificación y notificación radiológica asistida por ordenador.
Descripción del dispositivo	cmTriage de CureMetrix es un dispositivo que captura y marca FFDM con hallazgos sospechosos, las deposita en el sistema de archivo y comunicación de imágenes (PACS por sus siglas en inglés) y proporciona una notificación pasiva mediante un código cmTriage que indicará «Sospechoso» o « » (en blanco) al especialista en radiología, indicando la existencia de un caso que potencialmente puede beneficiarse de la priorización de ese especialista en radiología.	HealthMammo de Zebra es una herramienta que ayuda a priorizar y clasificar los hallazgos sospechosos en la exploración. Analiza automáticamente las mamografías de cribado 2D FFDM y marca y notifica la presencia de hallazgos sospechosos. No sustituye a la evaluación completa según el estándar de atención y no recomienda un tratamiento ni proporciona un diagnóstico.	CogNet QmTRIAGE es una herramienta que analiza mamografías de cribado 2D FFDM y notifica a un PACS la presencia de hallazgos sospechosos de cáncer en un estudio, lo que permite a los especialistas en radiología priorizar su lista de trabajo y les ayuda a visualizar los estudios priorizados. No sustituye a la evaluación completa según el estándar de cuidados.

Dispositivo (Compañía)	cmTriage (CureMetrix) (15)	HealthMammo (Zebra Medical Vision) (16)	CogNet QmTRIAGE (Med Cognetics) (17)
Indicaciones de uso	cmTriage es una herramienta de flujo de trabajo paralelo, de notificación pasiva para priorización exclusiva, que utilizan los especialistas en radiología para priorizar pacientes específicos dentro de la lista de trabajo de imágenes de atención estándar para mamografías de cribado 2D FFDM. cmTriage no envía una alerta proactiva directamente al especialista en radiología.	Zebra HealthMammo es una herramienta de flujo de trabajo paralelo de notificación pasiva que utilizan los médicos intérpretes cualificados por la MQSA para priorizar pacientes con hallazgos sospechosos a nivel de examen. Se limita a la categorización de exámenes, no proporciona ninguna información diagnóstica más allá del triaje y la priorización.	El software QmTRIAGE™ es una herramienta de flujo de trabajo paralelo de notificación pasiva utilizada por médicos intérpretes cualificados por la MQSA para priorizar pacientes con hallazgos sospechosos en el entorno de la atención médica. Se limita a la categorización de los exámenes, no proporciona ninguna información diagnóstica más allá del triaje y la priorización.
Usuarios	Especialistas en radiología.	Médico intérprete.	Médico intérprete.
Tipo de mamografía	DM.	DM.	DM.
Proveedor mamografía	Agnostic.	Hologic.	Hologic.
Análisis clínicos: diseño del estudio de lectores	Estudio multicéntrico, retrospectivo y ciego. La muestra fue enriquecida con cánceres confirmados por biopsia.	Estudio retrospectivo.	Estudio de cohortes retrospectivo.
Análisis clínico: número de casos analizados.	1.255 casos (400 cánceres).	835 casos (400 cánceres).	800 casos (399 cánceres).
Análisis clínico: resultados	AUC = 0,951 (0,937-0,964). Sensibilidad media = 86,9 % (83,6 %-90,2 %). Especificidad media = 88,5 % (83,6 %-90,2 %).	AUC = 0,966 (0,955-0,977). Sensibilidad (modo normal) = 89,89 % (86,69 %-92,38 %). Especificidad (modo normal) = 90,75 % (87,51 %-93,21 %).	AUC = 0,957 (0,936-0,974). Sensibilidad global = 87 %. Especificidad global = 89 %.

I.5. Justificación

Desde la década de los 90 en España en las distintas CC.AA. y ciudades autónomas se organizaron PDPCM dirigidos a mujeres asintomáticas con el fin de disminuir la mortalidad por cáncer de mama, mejorar el pronóstico y aumentar la calidad de vida de las mujeres afectadas. La detección precoz del cáncer de mama es la mejor opción para cumplir con los objetivos señalados, aunque se ha observado que la efectividad de los PDPCM puede estar limitada por la técnica de cribado utilizada, lo que puede conducir a una tasa sustancial de FP y FN, con el consiguiente sobrediagnóstico aparejado al cribado y tumores no diagnosticados durante el cribado ya sea por error o porque no son visibles para el especialista en radiología, como los cánceres de intervalo. Además, también se ha percibido un aumento importante en la carga de trabajo de los especialistas en radiología que leen las mamografías, derivada del elevado número de éstas, de la doble lectura y de la creciente falta de especialistas en radiología. El desarrollo de sistemas de IA basados en arquitecturas de redes neuronales convolucionales de aprendizaje profundo como herramientas de apoyo a la detección y diagnóstico radiológico y a la clasificación y notificación radiológica pueden ser una solución a los problemas mencionados de los PDPCM, pudiendo mejorar la detección de lesiones malignas, reducir los cánceres de intervalo, reducir la carga de trabajo de los especialistas en radiología e incluso mejorar la ratio entre coste y beneficio.

Este informe de evaluación de tecnologías sanitarias (ETS) tiene como destinatarios fundamentales las autoridades sanitarias nacionales y regionales del SNS español y pretende dar respuesta a la solicitud de la Comisión de prestaciones, aseguramiento y financiación (CPAF) en el proceso de identificación y priorización de necesidades de evaluación dentro del Plan de Trabajo Anual de la Red Española de Agencias de Evaluación de Tecnologías Sanitarias y prestaciones del SNS. El informe se realiza a petición de la Dirección General de Salud Pública y Equidad en Salud ante la necesidad de disponer de evidencia científica sobre la eficacia clínica, coste-efectividad e impacto presupuestario de la utilización de sistemas de IA en los PDPCM con el objeto de asesorar en la toma de decisiones sobre la posible incorporación de dichos sistemas en la cartera común básica del SNS.

II. Revisión de la evidencia

II.1. Objetivo

II.1.1. Objetivo general

Evaluar la eficacia clínica y la eficiencia de incorporar los sistemas de IA a los programas de detección precoz de cáncer de mama mediante una revisión sistemática (RS) de la evidencia científica.

II.1.2. Objetivo específico

Evaluar mediante una RS de la evidencia científica la precisión de la IA en la detección del cáncer de mama cuando se integra en los programas de detección precoz de cáncer de mama.

Evaluar mediante una RS de la evidencia científica el coste-efectividad de la incorporación de la IA en los programas de detección precoz de cáncer de mama.

II.2. Metodología

II.2.1. Revisión sobre la evidencia científica

Para evaluar la eficacia y la eficiencia de incorporar la IA en los PDPCM se realizó una RS de la evidencia científica con la finalidad de proveer de información objetiva que permitiese avalar la toma de decisiones en el cuidado en la salud, así como en las políticas sanitarias.

La metodología se basó en una búsqueda estructurada en bases de datos de literatura científica prefijadas, lectura crítica de los estudios, síntesis de los resultados y valoración de estos en relación con el contexto del SNS.

Como primer paso para sistematizar la búsqueda de bibliografía se transformaron los objetivos específicos en preguntas de investigación. Estas preguntas fueron planteadas primero en lenguaje natural y después en formato PICO (población, intervención, comparación, *outcomes*/resultados), para facilitar la identificación de los términos de búsqueda.

Pregunta de investigación 1. *¿El uso de sistemas de IA integrados en los PDPCM para la detección de cáncer de mama, es más, igual o menos preciso en comparación con la estrategia de cribado habitual realizada en los PDPCM?*

Pregunta de investigación formato PICO:

Pacientes: mujeres participantes en los PDPCM.

Intervención: sistemas de IA basados en arquitecturas de redes neuronales convolucionales de aprendizaje profundo empleados para la detección radiológica de cáncer de mama.

Comparación: cribado mamográfico habitual.

Medidas de resultado: área bajo la curva de la curva ROC (AUC-ROC, por sus siglas en inglés), sensibilidad y especificidad.

Método de abordaje

RS de la evidencia.

Pregunta de investigación 2. *¿Es coste-efectivo el uso de sistemas de IA en el cribado mamográfico de cáncer de mama en mujeres participantes en los PDPCM en comparación con la estrategia de cribado habitual realizada en los PDPCM?*

Pregunta de investigación formato PICO

Pacientes: mujeres participantes en los PDPCM.

Intervención: sistemas de IA basados en arquitecturas de redes neuronales convolucionales de aprendizaje profundo empleados para la detección radiológica de cáncer de mama.

Comparación: cribado mamográfico habitual.

Medidas de resultado: costes, años de vida ajustados por calidad (AVAC), años de vida ganados (AVG), efectividad incremental, coste incremental, ratio coste-efectividad incremental (RCEI), impacto presupuestario.

Método de abordaje

RS de la evidencia científica.

II.2.2. Fuentes de información y estrategia de búsqueda bibliográfica

Para dar respuesta a la primera pregunta de investigación, se llevó a cabo una búsqueda bibliográfica para la identificación de RS y/o metaanálisis (MA) en las siguientes bases de datos:

- Bases de datos de RS/MA/Informes: Cochrane Library (Wiley) e International HTA Database (Inahta).
- Bases de datos generales: Medline (PubMed) y Embase (OvidWeb).

También se realizaron búsquedas en bases de datos seleccionadas como Web of Science y Scopus. De igual forma, se procedió a la revisión manual de las referencias de los trabajos incluidos con la finalidad de localizar aquellos estudios que no se recuperaron en las búsquedas automatizadas.

La búsqueda fue ejecutada en julio de 2022. La estrategia incluía, entre otros, los siguientes términos en lenguaje libre y controlado: *breast neoplasm, artificial intelligence y screening*. Los estudios se limitaron a aquellos cuyo idioma de publicación fuera castellano o inglés.

A continuación, para la localización de estudios primarios originales se tomó como base la estrategia de búsqueda realizada en la revisión sistemática de Freeman *et al.* en 2021 (5). Esta búsqueda se actualizó solo a este tipo de publicaciones y se ejecutó en diciembre de 2022 en las siguientes bases de datos Medline (Pubmed), Embase (OvidWeb) y Cochrane.

Para dar respuesta a la segunda pregunta de investigación, se llevó a cabo una búsqueda bibliográfica para la identificación de estudios de costes y de evaluación económica en las principales bases de datos de estudios económicos: NHS EED (NHS Economic Evaluation Database) y CEA (Cost-Effectiveness Analysis) Registry, así como en las bases de datos Medline y Embase, mediante el empleo de términos y filtros económicos. Dicha búsqueda se realizó en diciembre de 2022 para un horizonte temporal de 10 años.

Se establecieron alertas en las bases de datos principales (Medline y Embase), con el objetivo de identificar estudios que se publicasen hasta la edición de este documento.

Las estrategias de búsqueda detalladas por cada base de datos se pueden consultar en el Anexo VI.1.

II.2.3. Criterios de selección de los estudios

Criterios de inclusión

Tipos de participantes

Mujeres participantes en PDPCM.

Prueba índice

Algoritmos de IA basados en arquitecturas de redes neuronales convolucionales de aprendizaje profundo aplicados a FFDM o a DBT de mujeres para la detección de cáncer de mama, ya sean como parte de un cambio de vía (como ayuda al especialista en radiología o como sistema autónomo que sustituye a alguno de los especialistas en radiología) o como una lectura completa (clasificación y notificación radiológica).

Estándar de referencia

Como estándar de referencia se adoptó el señalado en la RS realizada por Freeman *et al.* en 2021 (5):

Cáncer confirmado por análisis histológico de muestras de biopsia en mujeres remitidas para pruebas adicionales en el cribado y preferiblemente de presentación sintomática durante el seguimiento.

Los estudios tendrán una verificación diferencial ya que no todas las mujeres pueden o deben someterse a una biopsia. En estudios retrospectivos o de conjunto de pruebas enriquecidas (con lectores prospectivos) la decisión de si una mujer recibe biopsia o seguimiento se basa en la decisión del lector principal, lo que introduce sesgos, ya que el cáncer cuando esté presente tiene más probabilidades de detectarse si la persona se somete a pruebas de seguimiento después de la revaloración del cribado.

Cuando la IA se utiliza para la preselección para clasificar qué mamografías necesitan ser examinadas por un especialista en radiología y cuáles no, se acepta una definición de mamografía normal como aquella libre de cáncer detectado por cribado basada en la lectura humana consensuada.

Tipos de estudios

Los estudios deben evaluar la integración de la IA en PDPCM y no el desarrollo de un sistema de IA.

Los estudios deben informar sobre la precisión de las pruebas de algoritmos de IA aplicados a FFDM o a DBT para la detección del cáncer de mama como parte de un cambio de vía o lectura completa.

Los estudios sobre la precisión de las pruebas serán: RS, ensayos clínicos aleatorizados (ECAs), estudios prospectivos o retrospectivos con validación geográfica, estudios de cohortes comparativos o estudios multilector-multicaso con conjuntos de pruebas enriquecidos.

Los estudios sobre el coste-efectividad de los sistemas de IA serán: estudios completos de evaluación económica (EE) (coste-efectividad o coste-utilidad) y estudios de impacto presupuestario.

Fecha de publicación

Estudios publicados hasta diciembre de 2022.

Idioma de publicación

Estudios publicados en inglés o español.

Criterios de exclusión

Los criterios de exclusión se tomaron de la revisión sistemática realizada por Freeman *et al.* en 2021 (5):

- Estudios que informan de la validación de sistemas de IA mediante conjuntos de pruebas de validación interna (validación cruzada «x-fold», método «leave one out»), conjunto de pruebas de validación dividida, conjunto de pruebas de validación temporal.
- Estudios en los que < 90 % de las mamografías son FFDM o DBT.
- Estudios en los que la IA se utiliza para predecir el riesgo futuro de cáncer.
- Estudios en los que solo se informa de la detección de subtipos de cáncer.
- Estudios en los que se utiliza sistemas tradicionales de detección asistida por ordenador sin aprendizaje automático.
- Estudios en las que las medidas de precisión no se expresan en umbrales clínicamente relevantes (curva AUC-ROC), FP positivos y FN (por ejemplo, sensibilidad solo para muestras positivas de cáncer).
- Estudios en los que se obtienen resultados de simulación de la hipotética integración de la IA con las decisiones de los especialistas en radiología.

II.2.4. Proceso de selección de estudios

La selección de los estudios se realizó por pares. Para cada pregunta de investigación, dos revisores evaluaron de forma paralela e indepen-

diente los títulos y resúmenes de todos los estudios recuperados como potencialmente relevantes a través de la búsqueda de la literatura.

Aquellos estudios que cumplieron con los criterios de inclusión o aquellos en los que no hubo información suficiente para tomar una decisión clara fueron seleccionados para su lectura a texto completo. De forma independiente los dos revisores analizaron exhaustivamente los artículos a texto completo. Posteriormente, procedieron a la puesta en común de resultados y determinaron los estudios que finalmente se incluyeron para la síntesis de la evidencia. Cuando hubo duda y/o desacuerdo entre los revisores, esta se resolvió tras discusión y, cuando no hubo consenso, se consultó con otro revisor.

II.2.5. Extracción de datos y síntesis de la evidencia

De los artículos incluidos en esta RS un revisor extrajo los datos en un formulario de recogida de datos prediseñado. Las hojas de extracción de datos fueron comprobadas por un segundo revisor y los desacuerdos se resolvieron mediante discusión.

Para cada pregunta de investigación se diseñaron tablas en formato Word y Excel en las que se extrajeron las variables consideradas de interés para la RS de acuerdo con los objetivos generales y específicos señalados en el apartado correspondiente. De manera resumida, la información específica extraída para cada pregunta fue la siguiente:

Pregunta 1

- *Identificación del estudio*: autores, fecha de publicación, localización.
- *Diseño y metodología*: tipo de estudio, rol de la IA, población, proveedor mamografía, prueba índice, umbral prueba índice, comparador, umbral comparador, estándar de referencia.
- *Resultados*: tablas de contingencia 2x2, curva AUC-ROC, sensibilidad, especificidad, tiempo de lectura.
- *Conclusiones*.
- *Calidad de la evidencia*.

Pregunta 2

- *Identificación del estudio*: autores, fecha de publicación, localización.

- *Diseño y metodología*: tipo de evaluación, objetivo, población, intervención comparador, efectividad, costes, perspectiva, horizonte temporal, tasa de descuento, modelo, análisis de sensibilidad.
- *Resultados*: desenlaces de efectividad y de costes, desenlaces incrementales, RCEI, análisis de sensibilidad.
- *Conclusiones*.
- *Calidad de la evidencia*.

Para cada pregunta de investigación, se llevó a cabo un análisis descriptivo y narrativo de la evidencia y una síntesis de las principales medidas de resultado. La información se presentó cuantitativa y cualitativamente en función de la evidencia identificada. No se realizó un MA de los estudios evaluados debido a su escaso número y a su amplia heterogeneidad.

II.2.6. Evaluación de la calidad

La evaluación de la calidad metodológica de los estudios incluidos se realizó mediante el empleo de las siguientes herramientas:

- Para las RS se empleó la herramienta AMSTAR-2 (18). AMSTAR-2 permite evaluar tanto RS de ensayos aleatorizados como de estudios no aleatorizados. Es un cuestionario que contiene 16 dominios y aunque no proporciona una calificación global, de las debilidades en los siete dominios considerados críticos (protocolo registrado antes de la revisión, adecuada búsqueda de la literatura, justificación de los estudios excluidos, riesgo de sesgo de los estudios individuales incluidos, métodos de MA apropiados, consideración del riesgo de sesgo en la interpretación de los resultados de la revisión y evaluación de la presencia y el impacto probable del sesgo de publicación) surgen cuatro niveles de confianza: alta, moderada, baja y críticamente baja.
- Para los estudios primarios individuales se empleó la herramienta QUADAS-2 (19). QUADAS-2 evalúa el riesgo de sesgo de un estudio a través de cuatro dominios: selección de participantes, prueba índice, estándar de referencia y flujos y tiempos. Los tres primeros dominios también se evalúan para identificar problemas de aplicabilidad. Para cada dominio, los estudios se califican como «bajo», «alto» o «poco claro» en cuanto al riesgo de sesgo; y como «bajo», «alto» o «poco claro» en cuanto a la aplicabilidad. Cada dominio comprende una serie de preguntas orientadoras para facilitar el proceso de emitir un juicio sobre el riesgo de sesgo en cada ámbito que se responden con «Sí», «No» o «Poco claro», y están redactadas de tal forma que «Sí» indica un riesgo de sesgo bajo. En

el estudio se adaptó QUADAS-2, de acuerdo con lo propuesto en la revisión sistemática de Freeman *et al.* (5), a la pregunta de revisión específica modificando las preguntas de orientación en consecuencia y proporcionando una guía sobre cómo evaluar el riesgo de sesgo y las calificaciones de los problemas de aplicabilidad (Anexo VI.2.1).
- Para las EE se utilizaron las Fichas de Lectura Crítica (FLC 3.0) desarrollada por el Servicio de Evaluación de Tecnologías Sanitarias del País Vasco, Osteba (20). Las FLC 3.0 (www.lecturacritica.com) han sido diseñadas para realizar lectura crítica de distintos tipos de estudios epidemiológicos. Incluyen seis dominios que fueron valorados en cada estudio: pregunta de investigación, método, resultados, conclusiones, conflicto de intereses y validez externa. Los estudios se categorizaron como estudios de alta calidad, calidad media o baja calidad.

Para cada pregunta de investigación, la calidad metodológica de cada estudio fue evaluada y valorada por pares de revisores de forma independiente. En el caso de divergencias se reevaluaron los estudios y se llegó a un consenso.

II.3. Resultados

II.3.1. Resultados pregunta de investigación 1

¿El uso de sistemas de IA integrados en los PDPCM para la detección de cáncer de mama, es más, igual o menos preciso en comparación con la estrategia de cribado habitual realizada en los PDPCM?

II.3.1.1. Resultados de la búsqueda bibliográfica

RS y metaanálisis

La búsqueda bibliográfica realizada en las bases de datos electrónicas identificó 336 estudios como potencialmente relevantes. Una vez eliminadas las referencias duplicadas, se identificaron 181 para su lectura por título y resumen. Excluidos aquellos que no cumplieron con los criterios de inclusión, se seleccionaron 13 referencias para su lectura a texto completo. De estas se seleccionó una para el análisis de su calidad y síntesis de la evidencia.

Los estudios excluidos tras la lectura a texto completo y las razones de su exclusión se recogen en el Anexo VI.3.1.

En la figura 2 se muestra el diagrama de flujo que resume el proceso de selección de estudios para responder a la pregunta de investigación.

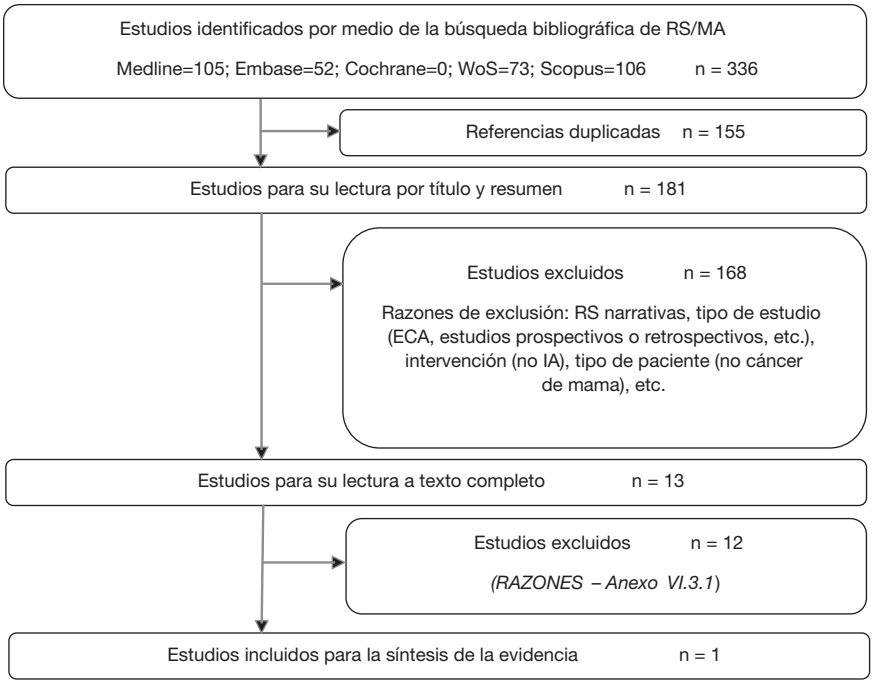


Figura 2. **Diagrama de flujo del proceso de selección de los estudios**

Estudios primarios

La búsqueda bibliográfica realizada en las bases de datos electrónicas identificó 2.232 estudios como potencialmente relevantes. Una vez eliminadas las referencias duplicadas, se identificaron 1.512 para su lectura por título y resumen. Excluidos aquellos que no cumplieron con los criterios de inclusión, se seleccionaron 31 referencias para su lectura a texto completo. De estas se seleccionaron 11 para el análisis de su calidad y síntesis de la evidencia.

Los estudios excluidos tras la lectura a texto completo y las razones de su exclusión se recogen en el Anexo VI.3.2.

En la figura 3 se muestra el diagrama de flujo que resume el proceso de selección de estudios para responder a la pregunta de investigación.

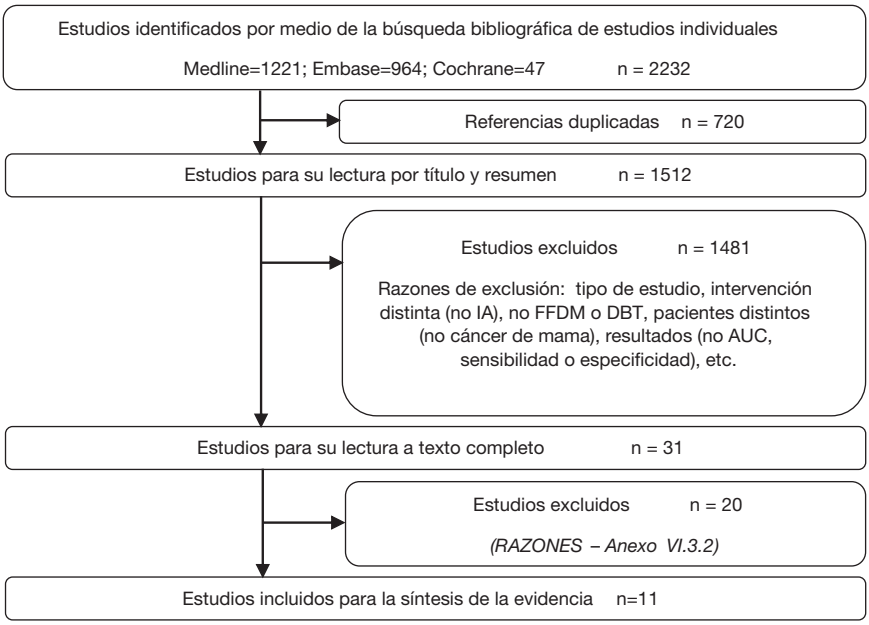


Figura 3. **Diagrama de flujo del proceso de selección de los estudios**

II.3.1.2. Descripción de los estudios incluidos

II.3.1.2.1. Características de la RS

En 2021 Freeman *et al.* (5) llevaron a cabo una RS con el objetivo de evaluar la precisión de la IA en la detección de cáncer de mama dentro de los PDPCM, poniendo especial atención en el tipo de cáncer detectado. Esta revisión se realizó en el Reino Unido (R.U.) y analizó el papel potencial que la IA podría jugar en las distintas fases del proceso de cribado de mama: como instrumento de apoyo para los especialistas en radiología, como lector autónomo o como herramienta de clasificación previa. Emplearon criterios de inclusión y exclusión precisos con el fin de centrar el análisis en los programas de cribado de cáncer de mama en lugar de en el desarrollo de sistemas de IA. Así, los estudios elegidos debían informar sobre la precisión de las pruebas de los algoritmos de IA aplicados a FFDM realizadas en mujeres para la detección de cáncer de mama en la práctica de cribado o en conjuntos de pruebas. Para ello seleccionaron estudios prospectivos, ECAs, estudios retrospectivos con validación geográfica, estudios de cohortes comparativos y estudios multilector-multicaso enriquecidos. El estándar de referencia que utilizaron en el análisis fue biopsia con histología o seguimiento en mujeres con cribado negativo. Como resultado

principal evaluaron la precisión de la prueba y como resultado secundario el tipo de cáncer detectado.

El resumen de las características principales de la RS incluida en el análisis queda reflejado en la tabla 3.

Tabla 3. **Características principales de la RS**

Autor	Objetivo	Rol de la IA	Criterios inclusión	Criterios exclusión	Tipo de estudios elegibles	Estándar de referencia	Resultados
Freeman, K. *et al*., 2021, R.U. (5)	Evaluar la precisión de la IA en la detección de cáncer de mama dentro de los PDPCM, centrándose en el tipo de cáncer detectado.	IA como sistema independiente. IA para clasificación. IA como ayuda al lector (en combinación con el especialista en radiología).	Estudios que informaron sobre la precisión de la prueba de los algoritmos de IA aplicados a las mamografías digitales de las mujeres para detectar el cáncer de mama, como parte de un cambio de ruta o una lectura completa (lectura + decisión que da como resultado la clasificación).	Estudios que informaron la validación de los sistemas de IA mediante conjuntos de pruebas de validación interna, mediante pruebas de validación divididas y pruebas de validación temporal. Estudios en los que menos del 90 % de las mamografías incluidas fueron FFDM de mamografías de cribado. Estudios en los que el sistema de IA se usó para predecir el riesgo futuro de cáncer o se informó sobre la detección de subtipos de cáncer. Estudios en los que se usaron sistemas tradicionales de detección asistida por ordenador sin aprendizaje profundo. Estudios en los que las medidas de precisión de la prueba no se expresaron en ningún umbral clínicamente relevante (p. ej., solo el AUC) o no caracterizó el equilibrio entre resultados FP y FN (p. ej., sensibilidad solo para muestras positivas para cáncer). Estudios que simulan la hipotética integración de la IA con las decisiones de los especialistas en radiología.	Estudios prospectivos, ECA, estudios retrospectivos validados en áreas geográficas, estudios comparativos de cohortes, estudios de laboratorio multilector-multicaso con pruebas enriquecidas.	Cáncer confirmado por análisis histológico de muestras de biopsia de mujeres remitidas para pruebas adicionales en el cribado y por presentación sintomática durante el seguimiento.	Resultado principal: precisión de la prueba. Resultado secundario: tipo de cáncer y cánceres de intervalo.

II.3.1.2.2. Características de los estudios individuales

Las características principales de los 11 estudios incluidos (21-31) en esta RS para su análisis quedan reflejadas en la tabla 4. Todos los estudios fueron publicados en 2021 y 2022, ocho fueron estudios retrospectivos (21, 23, 26-31), dos estudios multilector-multicaso (22, 25) y uno retrospectivo para determinar la efectividad del modelo y prospectivo para investigar la aplicación clínica del modelo (24). En nueve estudios se analizaron FFDM (21-24, 26, 27, 29-31,), en uno DBT (25) y en otro FFDM y DBT (28).

El objetivo general de los estudios fue evaluar la precisión de la IA en la detección de cáncer de mama en el cribado de mama. Los datos necesarios para la realización de los análisis propuestos en los diferentes estudios se obtuvieron de conjuntos de datos procedentes de cohortes poblacionales de China (21, 24), EE. UU. (25, 26), Francia (22), Corea del Sur (23), Dinamarca (31), Alemania (27), España (28), Noruega (30) y R.U. y Hungría (29). En tres estudios (22, 25, 27) se señaló que las cohortes fueron enriquecidas con casos de cáncer, en cuatro (21, 23, 24, 31) se indicó que las cohortes presentaron una alta proporción de casos de cáncer lo que no reflejaba el cribado de cáncer de mama en la práctica clínica, mientras que en otros cuatro (26, 28, 29, 30) las cohortes de cribado seleccionadas fueron representativas del cribado de la vida real. La población incluida fue de mujeres participantes en programas de cribado de cáncer de mama (22, 26-31) de entre 50 y 69-70 años y de mujeres a las que se realizó cribado de mama (21, 23-25) empleando como técnicas de cribado FFDM o DBT.

En todos los estudios, los sistemas de IA analizados utilizaron redes neuronales convolucionales de aprendizaje profundo. En ocho (21-23, 25, 28-31) se emplearon sistemas de IA comercialmente disponibles (Transpara 1.7.0 o 1.6.0, Mammoscreen v1.2., Lunit v1.1.1.0, Mia v2.0.1 y Yizhun 3.2.3) y en tres (24, 26, 27) sistemas propios (*in-house*). En siete estudios (22, 23, 25, 26, 28, 30, 31) los sistemas de IA utilizaron puntuaciones bien para indicar la probabilidad de malignidad (puntuación entre 1 y 100 (100 máxima sospecha)) o bien para señalar la probabilidad de cáncer visible en la mamografía (puntuación entre 1 y 10 (10 máxima sospecha) o entre 0 y 1 (1 máxima sospecha)); en dos estudios (21, 24) los sistemas de IA representaron el grado de malignidad utilizando puntuaciones del sistema de datos e informes de imágenes mamarias (BI-RADS por sus siglas en inglés); en uno (27) el sistema de IA evaluó la confianza de la capacidad predictiva de la IA: confía (cribado normal/cáncer), no confía; y en uno último (29) el sistema realizó una sugerencia binaria: «volver a llamar» indicando sospecha de malignidad, «no volver a llamar» hasta el siguiente intervalo de cribado.

En cinco estudios (21-25), los sistemas de IA se evaluaron como ayuda al lector en el proceso de cribado de mama; en tres (26, 28, 29), como sistema autónomo en sustitución de uno de los lectores en la doble lectura; en dos (30, 31) para clasificación de mamografías como normales o sospechosas de malignidad; y en uno (27), como sistema autónomo y para clasificación. En los estudios en los que los sistemas de IA se utilizaron como sistemas autónomos se analizaron un total de 496.503 FFDM (6.421 casos de cáncer, prevalencia: 1,29 %, rango: 0,59-3,37 %) y de 15.999 DBT (133 casos de cáncer, prevalencia: 0,71 %). En los que los sistemas de IA se emplearon como ayuda al lector se analizaron un total de 1.588 FFDM (993 casos de cáncer, prevalencia: 62,53 %, rango: 20,38-73,77 %) y de 480 DBT (146 casos de cáncer, prevalencia: 30,42 %). Por último, en los estudios en los que los sistemas de IA se usaron para clasificación el número total de FFMD analizadas fue de 320.241 (4.868 casos de cáncer, prevalencia: 1,52 %, rango: 0,78-3,37 %).

En los sistemas de IA empleados como ayuda al lector el rendimiento de los algoritmos se comparó frente a la lectura única realizada por especialistas en radiología sin ayuda de la IA. En tres estudios (22-24) 12, 10 y 12 especialistas en radiología, respectivamente, con diferente experiencia y certificación por la MQSA, leyeron FFDM en dos sesiones, con y sin ayuda de la IA, espaciadas entre ellas por un periodo de espera de cuatro semanas (22, 24) o de dos meses (23). En un estudio (21) 71 especialistas en radiología con diferente experiencia y formación leyeron las mamografías FFDM divididas aleatoriamente en dos grupos, 35 con ayuda de la IA y 36 sin ayuda. Por último, en otro estudio (25) 18 especialistas en radiología con certificado MQSA, con una experiencia media de nueve años y activos en la lectura de exámenes de DBT en la práctica clínica, interpretaron las mamografías DBT en dos sesiones, con y sin ayuda de la IA, con un tiempo de descanso entre sesiones de cuatro semanas. Además, nueve especialistas en radiología leyeron los exámenes con acceso a mamografía sintética y apoyo interactivo a la navegación. En los estudios en los que los sistemas de IA se utilizaron como sistemas autónomos, el rendimiento de estos se comparó frente a lectura única radiológica en un estudio (26), con lectura doble más consenso en dos estudios (27, 29) y con lectura única y lectura doble en un estudio (28). Por último, en los estudios en los que se utilizaron los sistemas de IA para la clasificación de mamografías (30, 31), su rendimiento se comparó con lectura doble más consenso.

Con respecto al estándar de referencia, cabe señalar que en los artículos incluidos los casos positivos de cáncer se confirmaron mediante biopsia y seguimiento que pudo variar entre seis meses (23, 25, 30 31), un año (26), dos años (24, 28), tres años (29) o no seguimiento (21, 22, 27). Además, dos

artículos (23, 24) definieron caso benigno como el confirmado mediante biopsia o imagen de seguimiento durante dos años o más desde el cribado, y tres artículos (29-31) definieron cáncer de intervalo como el diagnosticado a los dos (30, 31) o tres años (29) siguientes a un análisis de cribado negativo o a los dos años después de una revaloración con resultado negativo (30, 31). Los casos negativos se confirmaron mediante lectura normal y seguimiento negativo, seguimiento que pudo variar entre un año (25, 26), dos años (23, 24, 27, 28, 31), casi tres años (1.035 días) (29) o no seguimiento (22, 30).

Tabla 4. **Características de los estudios individuales**

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Bao *et al.* 2022 (21)	Estudio retrospectivo.	IA como ayuda al lector.	643 FFDM de campo completo obtenidas de dos hospitales 3A y uno terciario en Beijing, China. Mujeres, edad media 54 años.	GE, Philips, Hologic, Siemens, etc.	Sistema Yizhun versión 3.2.3. Umbral no especificado.	36 especialistas en radiología sin ayuda de IA. Lectura única. BI-RADS 1-3: benignidad. BI-RADS 4-5: malignidad. Umbral no especificado.	Cáncer: biopsia positiva.
Dang *et al*. 2022 (22)	Estudios multilector-multicaso con diseño transversal.	IA como ayuda al lector.	314 exámenes obtenidos de un hospital en Francia. Mujeres que participan en el programa de cribado de cáncer de mama. Edad entre 50-74 años, sin historia personal o familiar de cáncer de mama o de cirugía de mama y sin factores genéticos de riesgo.	Hologic Selenia 3D Dimension.	Sistema Mammoscreen ™ v.1.2 (Therapixel). Puntúa entre 1-10 de acuerdo con el nivel de sospecha de presencia de cáncer (1 benigno, 10 sospechoso). Umbral no especificado.	12 especialistas en radiología sin ayuda de IA. Lectura única. Utilizan puntuaciones «BI-RADS 100 continuo» definido con base a una escala de nivel de sospecha (NDS) entre 1-100: BI-RADS 1: NDS entre 1-20. BI-RADS 2: NDS entre 21-40. BI-RADS 3: NDS entre 41-60. BI-RADS 4: NDS entre 61-80 BIRADS 5: NDS entre 81-100. Umbral no especificado.	Cáncer: confirmado con biopsia positiva. No cáncer: verificado por seguimiento negativo.

…/…

…/…

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Lee *et al*. 2022 (23)	Estudio retrospectivo.	IA como ayuda al lector.	200 casos obtenidos del Hospital Universitario Soonchunhyang de Bucheon.	GE Healthcare.	Sistema Lunit INSIGHT MMG version 1.1.1.0. Puntúa entre 1-100 para indicar la probabilidad de malignidad (PDM). PDM ≥ 3, lesión sospechosa de cáncer de mama.	10 especialistas en radiología sin ayuda de IA. Lectura única. Utilizan una escala de 7 puntos para indicar la PDM de las mamografías: PDM 1: definitivamente normal. PDM 2 = benigna. PDM 3 = probablemente benigna. PDM 4 = baja sospecha de malignidad. PDM 5 = sospecha moderada de malignidad. PDM 6 = sospecha alta de malignidad. PDM 7 = altamente sugestivo de malignidad. PDM ≥ 3, lesión sospechosa de cáncer de mama.	Casos malignos: patológicamente confirmados con biopsia en los seis meses siguientes a la mamografía. Casos negativos: mamografías con BI-RADS categoría 1, confirmadas como negativo durante más de dos años de seguimiento. Casos benignos: confirmados por biopsia o imagen de seguimiento durante más de dos años.

…/…

.../...

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Sun *et al*. 2021 (24)	Estudio multicéntrico que incluye un estudio retrospectivo y uno prospectivo.	IA como ayuda al lector.	Estudio retrospectivo: 200 mujeres (70 con cáncer) a las que se realiza FFDM con diagnóstico patológico o más de dos años de seguimiento después del primer examen. Estudio prospectivo. 5.746 mujeres (832 con cáncer) a las que se realiza FFDM. Datos obtenidos de seis centros de Pekín, China.	GE Medical Systems Senographe, Siemens Mammomat Novation DR e Insiration, Hologic Selenia Dimensions, Philips.	Sistema de IA (*in-house*). Utiliza cinco modelos logit (modelos de elección binaria) para predecir el BI-RADS de cada lesión: Logit 1: BI-RADS de la lesión > que BI-RADS 3. Logit 2: BI-RADS de la lesión > que BI-RADS 4A. Logit 3: BI-RADS de la lesión > que BI-RADS 4B. Logit 4: BI-RADS de la lesión > que BI-RADS 4C. Logit 5: BI-RADS de la lesión > que BI-RADS 5. Umbral no especificado.	12 especialistas en radiología sin ayuda de IA. Lectura única. Emplean la clasificación BI-RADS (entre 1-5) para etiquetar pacientes con lesiones malignas o benignas y sin lesiones. Umbral no especificado.	Lesiones malignas: diagnóstico patológico en un plazo de dos años desde primera mamografía. Lesiones benignas: diagnóstico patológico benigno en un plazo de dos años; y benigna, sin diagnóstico patológico, después de más de dos años desde la primera mamografía.
Van Winkel *et al*. 2022 (25)	Estudios multilector-multicaso totalmente aleatorizado.	IA como ayuda al lector.	360 casos representativos de mujeres sometidas a exámenes de DBT de cribado y diagnóstico (214 cánceres) obtenidos de siete centros clínicos en EE.UU. Edad media 56,3 años.	Mammomat Inspiration (Siemens Helthineers).	Sistema de IA Transpara 1.6.0. (ScreenPoint Medical BV)). Utiliza puntuaciones entre 1-10 para indicar la probabilidad incremental de que un cáncer visible esté presente en la mamografía y puntuaciones entre 1-100: para indicar el nivel de sospecha para cáncer. Umbral no especificado.	18 especialistas en radiología certificados por la MQSA sin apoyo de IA. Lectura única. Umbral no especificado.	Cáncer: comprobado mediante biopsia o al menos seis meses de seguimiento. No cáncer: casos negativos con al menos un año de seguimiento.

.../...

…/…

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Hsu *et al*. 2022 (26)	Estudio retrospectivo.	IA como sistema independiente (autónomo).	41.343 mujeres, 121.753 exámenes de FFDM, extraídos de la cohorte 2D UCLA Health Athena. Mujeres que acuden a un centro ambulatorio para someterse a examen mamográfico de cribado. Se incluyen diagnósticos de cáncer de mujeres que pueden haber recibido biopsias o diagnósticos fuera de la UCLA.	Hologic.	Dos sistemas de IA (*in-house*): un modelo conjunto de desafío (CEM) de los 11 modelos aportados por los seis mejores equipos de la fase competitiva del DREAM Mammography Challenge y un modelo CEM más la puntuación BI-RADS global proporcionada por el especialista en radiología intérprete original a nivel de examen (CEM+R). Umbral: sensibilidad (82,6 %) o especificidad (93 %) del especialista en radiología estimada a partir de la misma muestra	Especialista en radiología único. Puntuaciones BI-RADS 1 y 2 sospecha baja de malignidad. Puntuaciones BI-RADS 0, 3, 4 y 5 sospecha alta de malignidad.	Cáncer: exámenes de cribado con diagnóstico de cáncer (biopsia) en los 12 meses siguientes. No cáncer: exámenes de cribado sin diagnóstico de cáncer en los 12 meses siguientes, o con al menos 12 meses de diferencia sin diagnóstico de cáncer en ese periodo.
Leibig *et al*. 2022 (27)	Estudio de análisis retrospectivo. Los estudios sospechosos que se presentaron a consenso se enriquecieron durante la recogida de datos.	IA como sistema independiente (autónomo). IA para clasificación (triaje).	1.193.197 FFDM (453.104 mujeres) extraídas de ocho centros de cribado en Alemania. Mujeres asintomáticas con una categoría de densidad mamaria B o C de acuerdo con la ACR participantes cada 2 años en el programa nacional de cribado de cáncer de mama. Edad: 50-70 años.	Siemens, Hologic, Fuji y otros.	Sistema de IA (*in-house*). Umbral IA como sistema independiente: sensibilidad del especialista en radiología (86 %). Umbral IA para clasificación: se establecen dos umbrales para categorizar la derivación de decisiones: triaje normal, red de seguridad y derivación al especialista en radiología. Los umbrales se representan como conjuntos de dos puntos operativos: triaje normal (sensibilidad del algoritmo en el conjunto de datos de validación) + red de seguridad (sensibilidad del algoritmo en el conjunto de datos de validación).	Lectura doble más consenso. Puntuación BI-RADS 1 y 2, no revaloración. Puntuación BI-RADS ≥ 3, revaloración.	Cáncer: cáncer positivo confirmación histopatológica (biopsia). No cáncer: mamografías normales en un plazo mínimo de 24 meses, que no se habían revalorado (BI-RADS 1 o 2) o, en el caso de un hallazgo, el estudio de seguimiento debía haberse considerado negativo mediante doble lectura, conferencia de consenso o revaloración negativa.

…/…

.../...

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Romero-Martin *et al*. 2022 (28)	Estudio de cohortes con datos retrospectivos (los datos se recogieron del ensayo de cribado con tomosíntesis de Córdoba).	IA como sistema independiente (autónomo).	15.999 exámenes de un centro en Córdoba (España) obtenidos de 15.998 (113 cánceres) mujeres participantes en el programa de cribado de cáncer de mama. Edad 50-69 años (edad media 58 años).	Selenia Dimensions (Hologic).	Sistema de IA Transpara 1.7.0. (ScreenPoint Medical). Una puntuación entre 1-100 indica la probabilidad incremental de que un cáncer visible esté presente en la mamografía. Umbral: se identifican cuatro puntos operativos de la IA que producen una sensibilidad no inferior en comparación con los escenarios de cribado originales: 80 para lectura única de DM. 74 para lectura doble de DM. 65 para lectura única de BDT. 57 para lectura doble de BDT.	Cuatro escenarios de cribado originales con 4 especialistas en radiología: lectura única de DM, lectura doble de DM, lectura única de DBT y lectura doble de DBT. Umbral no especificado.	Cáncer: hallazgos histopatológicos de biopsia en los 24 meses siguientes al cribado. No cáncer: lectura normal en los dos años de seguimiento.

.../...

.../...

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Sharma *et al.* 2021 (29)	Cohorte histórica (retrospectiva) consecutiva de pacientes, no enriquecida.	IA como sistema independiente (autónomo).	La muestra incluyó 275.900 casos (177.882 participantes representativas de una población de cribado real) de siete centros de cribado, en los que participaron dos países (R.U. y Hungría). Mujeres entre 50-70 años invitadas a participar en el programa de cribado de mama del R.U. (intervalo de cribado de 3 años) y entre 45-65 años invitadas a participar en el programa de cribado de mama de Hungría (intervalo de cribado dos años). Se incluyeron mujeres fuera de los rangos de edad señalados.	Hologic, GE Healthcare, Siemens Healthneers y IMS Giotto.	Sistema de IA Mia ™ versión 2.0.1 «AI system (Kheiron Medical Technologies). Umbrales preespecificados para revalorar o no. No señalados.	Opinión aislada de un primer lector. Un segundo lector, a su discreción, tiene acceso a la opinión del primer lector. En caso de desacuerdo, arbitraje. Umbral: decisión de revalorar o no. No se especifica.	Cáncer detectado por el cribado: casos correctamente identificados por el flujo de trabajo histórico de doble lector, con una neoplasia maligna demostrada por patología confirmada por biopsia. Cáncer de intervalo: cribado con un cáncer con patología probada surgido en los tres años siguientes a la fecha original de cribado. No cáncer: caso de cribado con evidencia de un resultado de seguimiento negativo que incluyera una lectura de mamografía al menos 1.035 días después de la fecha de cribado original, sin prueba de malignidad entre medias.

.../...

…/…

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Larsen *et al*. 2022 (30)	Estudio retrospectivo.	IA para clasificación (triaje).	47.877 mujeres cribadas (122.969 exámenes de cribado, 957 cánceres) en cuatro unidades de cribado del programa poblacional de cribado de mama noruego (bienal). Mujeres entre 50 y 69 años.	Mammomat (Siemens Healthcare).	Sistema de IA Transpara 1.7.0. Se establecen tres umbrales: Umbral 1: puntuación bruta > 9 (puntuación global de 10), «seleccionado» por el sistema de IA; puntuación < 10 «no seleccionado». Umbral 2: tasa de selección = 8,8 % (tasa de consenso) (puntuación bruta > 9,13). Umbral 3: tasa de selección = 5,8 % (puntuación bruta > 9,43).	Doble lectura. Umbral: se establece una puntuación entre 1 y 5 para indicar la sospecha de malignidad: 1 negativo para malignidad. 2 probablemente benigno. 3 sospecha de malignidad intermedia. 4 probablemente maligno. 5 sospecha alta de malignidad). Puntuación de interpretación ≥ 2 por cada especialista en radiología, consenso para determina si revalorar o no.	El cáncer detectado por cribado: cáncer de mama diagnosticado después de una revisión y en los seis meses siguientes al examen de cribado. Cáncer de intervalo: cáncer de mama diagnosticado en los 24 meses siguientes a un examen de cribado negativo o 6-24 meses después de revaloración con resultado negativo. No cáncer: evaluación negativa por parte de ambos especialistas en radiología, tras el consenso o tenían una revaloración con un resultado negativo.

…/…

…/…

Estudio	Diseño estudio	Rol de la IA	Población	Proveedor mamografía	Prueba índice	Comparador	Estándar de referencia
Lauritzen *et al.* 2022 (31)	Estudio retrospectivo.	IA para clasificación (triaje).	114.421 mujeres con FFDM, extraídas de la región Capital de Dinamarca. Mujeres asintomáticas, entre 50-69 años, que participan en el cribado bianual de cáncer de mama.	Mammomat Inspiration (Siemens Healthineers).	Sistema de IA Transpara 1.7.0. (ScreenPoint Medical). Umbral IA para clasificación: Puntuación de 0-10, que indica el riesgo de presencia de posible cáncer visible: umbral de exclusión (puntuación de 5) y umbral de revaloración (UR) (igual a 9,989) para categorizar qué mamografías son normales, de riesgo moderado y sospechosas: Puntuación < 5: mamografía normal. Puntuación ≥ 5 y ≤ 9,989: riesgo moderado (la leen los dos especialistas en radiología). Puntuación > 9,989: sospechosa (se revalora directamente).	Las mamografías son leídas de forma independiente por dos especialistas en radiología especializados. Si no hay acuerdo sobre la decisión revalorar, consenso con un tercer especialista en radiología. Umbral no especificado.	Cáncer detectado por el cribado: mujeres con un cribado positivo a las que se diagnosticó cáncer de mama o carcinoma ductal *in situ* en los seis meses siguientes al cribado. Cánceres de intervalo: mujeres con un cribado negativo o un examen de revaloración negativo diagnosticados en los 24 meses siguientes al cribado (o antes del siguiente cribado). No cáncer: mujeres sin hallazgos sospechosos (resultado negativo del cribado) dentro de los dos años siguientes al cribado.

II.3.1.3. Calidad de la evidencia de los estudios incluidos

II.3.1.3.1. Calidad de la evidencia de la revisión sistemática

En la tabla 5 queda reflejada la calidad metodológica de la RS incluida, calidad medida mediante los criterios AMSTAR-2.

Tabla 5. **Calidad de la RS incluida. Escala AMSTAR 2**

	Freeman, K. *et al.* (5)
1. ¿Las preguntas de investigación y los criterios de inclusión para la revisión incluyen los componentes de PICO?	Sí
2. ¿El reporte de la revisión contiene una declaración explícita de que los métodos de revisión fueron establecidos con anterioridad a su realización y justifica cualquier desviación significativa del protocolo?	Sí
3. ¿Los autores de la revisión explicaron su decisión sobre los diseños de estudio a incluir en la revisión?	Sí
4. ¿Los autores de la revisión usaron una estrategia de búsqueda bibliográfica exhaustiva?	Sí
5. ¿Los autores de la revisión realizaron la selección de estudios por duplicado?	Sí
6. ¿Los autores de la revisión realizaron la extracción de datos por duplicado?	Sí
7. ¿Los autores de la revisión proporcionaron una lista de estudios excluidos y justificaron las exclusiones?	Sí
8. ¿Los autores de la revisión describieron los estudios incluidos con suficiente detalle?	Sí
9. ¿Los autores de la revisión usaron una técnica satisfactoria para evaluar el riesgo de sesgo de los estudios individuales incluidos en la revisión?	Sí
10. ¿Los autores de la revisión reportaron las fuentes de financiación de los estudios incluidos en la revisión?	No
11. Si se realizó un metaanálisis, ¿los autores de la revisión usaron métodos apropiados para la combinación estadística de resultados?	N/A
12. Si se realizó un metaanálisis, ¿los autores de la revisión evaluaron el impacto potencial del riesgo de sesgo en estudios individuales sobre los resultados del metaanálisis u otra síntesis de evidencia?	N/A
13. ¿Los autores de la revisión consideraron el riesgo de sesgo de los estudios individuales al interpretar / discutir los resultados de la evaluación?	Sí

	Freeman, K. *et al.* (5)
14. ¿Los autores de la revisión proporcionaron una explicación satisfactoria y discutieron cualquier heterogeneidad observada en los resultados de la revisión?	Sí
15. Si se realizó una síntesis cuantitativa, ¿los autores de la revisión llevaron a cabo una adecuada investigación del sesgo de publicación (sesgo de estudio pequeño) y discutieron su posible impacto en los resultados de la revisión?	N/A
16. ¿Los autores de la revisión informaron de cualquier fuente potencial de conflicto de intereses, incluyendo cualquier financiamiento recibido para llevar a cabo la revisión?	Sí

Con base en el cumplimiento de los criterios AMSTAR-2, la RS incluida (5) se valoró con calidad metodológica alta al no mostrar debilidades en alguno de los siete dominios considerados críticos por esta herramienta.

II.3.1.3.2. Calidad de la evidencia de los estudios individuales

La calidad de los estudios primarios de precisión diagnóstica se valoró utilizando la herramienta QUADAS-2. Los resultados de esta evaluación quedan recogidos en la tabla 6 y en la tabla del Anexo VI.2.2. Un resumen del riesgo de sesgo y de los problemas de aplicabilidad de los estudios queda reflejado en la figura 4.

El riesgo de que la selección de pacientes (QUADAS-2, dominio 1) hubiera introducido sesgos se consideró bajo en cinco estudios (26, 28-31) y alto en seis (21-25, 27). Que el riesgo de sesgo fuese alto en estos estudios se debió a que emplearon para su análisis muestras enriquecidas de pacientes, lo que ocasionó una prevalencia de cáncer de mama atípica entre el 3,37 % y el 70 %, para la población de cribado. El empleo de muestras enriquecidas supuso, además, la observancia de problemas de aplicabilidad altos para este dominio. Asimismo, en tres de ellos (21, 23, 24) las mamografías analizadas fueron de población asiática con un elevado número de mujeres con alta densidad mamaria, no representativa de la población española que participa en los cribados.

Con respecto a la evaluación de la prueba índice (QUADAS-2, dominio 2) se observó riesgo de sesgo bajo en cinco estudios (26-28, 30, 31) y alto en seis (21-25, 29). La existencia de riesgo de sesgo alto en estos estudios se debió a que en ninguno se señaló un umbral de positividad para la prueba índice. Los resultados positivos y negativos de las pruebas en

los estudios sobre precisión se definen en función del nivel de positividad de la prueba y cambian si se altera el umbral. Esta dependencia del umbral es un aspecto fundamental de la evaluación de la precisión de las pruebas ya que induce un equilibrio entre sensibilidad y especificidad, una modificación del umbral de positividad aumenta un valor mientras el otro disminuye. Además, para este dominio los problemas de aplicabilidad se valoraron altos en todos los estudios por no haberse identificado un umbral de positividad en los estudios señalados anteriormente, porque en tres estudios (24, 26, 27) el sistema de IA no estuvo disponible comercialmente al ser sistemas *in-house* y porque en 10 estudios (21-23, 25-31) no se describió la precisión de los sistemas de IA integrada en una vía clínica de cribado de mama ni se evaluó la precisión de la IA de forma prospectiva.

El riesgo de sesgo observado para el patrón de referencia (QUADAS-2, dominio 3) se valoró bajo en siete estudios (23, 24, 27-31) y alto en cuatro (21, 22, 25, 26). El riesgo de sesgo se estimó alto porque el periodo de seguimiento de las mujeres con resultado negativo en el cribado o no se señaló o fue menor de dos años, lo que pudo dar lugar a una subestimación del número de cánceres no detectados y a una sobreestimación de la precisión de la prueba, tal y como señalaron Freeman *et al.* (5). Además, esto ocasionó problemas altos de aplicabilidad en los mismos estudios.

Por último, el riesgo de sesgo del dominio flujo y tiempo (QUADAS-2, dominio 4) se estimó poco claro en cinco estudios (21-25) y alto en seis (26-31). En los estudios retrospectivos, la decisión de revalorar para realizar biopsia o pruebas de seguimiento se tomó en función de la decisión de los especialistas en radiología originales y no con base en la decisión del sistema de IA por lo que no se supo si las mujeres positivas detectadas por la IA y las negativas detectadas por los especialistas en radiología fueron FP o verdaderos negativos (VN), ni el tipo de cáncer detectado como verdadero positivo (VP). El seguimiento hasta el desarrollo de los cánceres de intervalo pudo detectar alguno, pero no todos estos cánceres, lo que redujo, pero no evitó estos sesgos.

Tabla 6. **Tabla resumen del riesgo de sesgo y de los problemas de aplicabilidad: juicio de los autores de la revisión sobre cada dominio para cada estudio incluido**

	Riesgo de sesgo				Problemas de aplicabilidad		
	Selección de pacientes	Prueba índice	Patrón de referencia	Flujo y tiempo	Selección de pacientes	Prueba índice	Patrón de referencia
Bao 2022 (21)	+	+	+	?	+	+	+
Dang 2022 (22)	+	+	+	?	+	+	+
Lee 2022 (23)	+	+	−	?	+	+	−
Sun 2021 (24)	+	+	−	?	+	+	−
vanWinkel 2022 (25)	+	+	+	?	+	+	+
Hsu 2022 (26)	−	−	+	+	?	+	+
Leibig 2022 (27)	+	−	−	+	+	+	−
Romero-Martín 2022 (28)	−	−	−	+	−	+	−
Sharma 2021 (29)	−	+	−	+	−	+	−
Larsen 2022 (30)	−	−	−	+	−	+	−
Lauritzen 2022 (31)	−	−	−	+	−	+	−

+ Alto ? Poco claro − Bajo

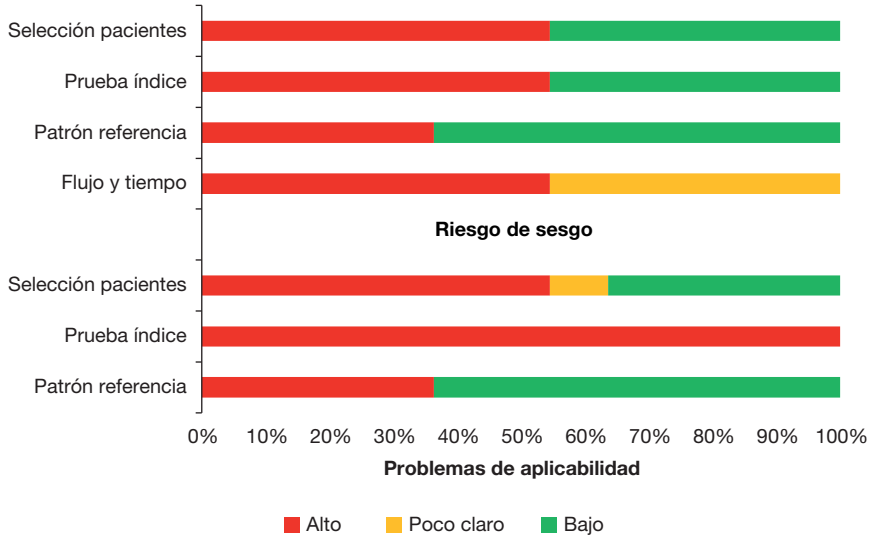


Figura 4. **Figura resumen del riesgo de sesgo y de los problemas de aplicabilidad. Los juicios de los autores de la revisión sobre cada dominio se presentan en forma de porcentaje en los estudios incluidos**

II.3.1.4. Descripción y análisis de los resultados

II.3.1.4.1. Descripción y análisis de los resultados de la revisión sistemática

Los resultados de la RS analizada en este informe quedan descritos en las tablas 7 y 8. En ambas tablas los resultados se muestran para los distintos papeles potenciales que la IA puede jugar en el proceso de cribado de cáncer de mama: como lector autónomo en sustitución de alguno de los especialistas en radiología, como instrumento de apoyo a los especialistas en radiología y como herramienta de clasificación previa al cribado.

IA como sistema autónomo para sustituir alguno de los especialistas en radiología

La capacidad discriminativa de los sistemas de IA para detectar pacientes con cáncer frente a pacientes sin cáncer se estimó mediante el AUC-ROC. En los cinco estudios (32-36) en los que la IA se analizó como un sistema autónomo, el parámetro AUC-ROC proporcionó unos valores que oscilaron entre 0,840 y 0,956. La capacidad discriminatoria de los sistemas de IA se consideró buena en tres estudios, AUC-ROC igual a

0,923 (36), a 0,956 (0,948-0,966) (35) y a 0,945 (0,919-0,968 (32); y aceptable en dos estudios, AUC-ROC igual a 0,889 (0,871-0,907) (33) y a 0,840 (0,820-0,860) (34). Solo en un estudio (34) se comparó el AUC-ROC de la IA (0,840) con el de los especialistas en radiología (0,814), siendo la diferencia estadísticamente no inferior, 0,026 (–0,003-0,055).

En dos estudios de pruebas enriquecidas multilector-multicaso (32, 34) se señaló que la sensibilidad y especificidad de un sistema de IA comercial (Transpara v1.4.0) (34) y de uno propio (32) fue superior a la sensibilidad y especificidad media de la lectura única de los especialistas en radiología en un entorno de laboratorio. Rodríguez-Ruíz *et al.* (34) indicaron que para la base de datos C (país de lectura Holanda) y para un umbral de predicción de la IA igual a la especificidad media de los especialistas en radiología (79 % (73-86 %)), la sensibilidad para el sistema de IA fue mayor que la sensibilidad media de los especialistas en radiología (80 % (70-90 %) vs 77 % (70-83 %), diferencia igual al 3 % (–6,2-12,6 %)). Por su parte, Lotter *et al.* (32) para el marco temporal que denominan exámenes de cáncer «índice», utilizados normalmente en los estudios de lectores y que eran las mamografías de cribado adquiridas más recientemente antes de la malignidad demostrada por biopsia, obtuvieron, para un umbral de predicción de la IA igual a la especificidad del lector (66,9 %), una sensibilidad del 96,2 % (91,7-99,2 %) superior ($P < 0,001$) a la de la lectura única (82 %) y, para un umbral de predicción igual a la sensibilidad del lector (82 %), una especificidad del 90,9 % (84,9-96,1 %) también superior ($P < 0,001$) a la de la lectura única (66,9 %). Ahora bien, dado que estos estudios se realizaron en un ambiente de laboratorio la generalización de estos resultados en la práctica clínica resulta incierta.

En otro estudio (33), en el que se empleó un sistema de IA entrenado exclusivamente con datos de R.U. y comprobado con un conjunto de datos de EE.UU. (3.097 mujeres, 22,2 % cáncer) para generalizar la IA entre la población y los sistemas de salud, también se observó que el sistema de IA superó en sensibilidad (56,24 % vs 48,10 %, $P = 0,0006$) y especificidad (84,29 % vs 80,83 %, $P = 0,0212$) a la lectura única realizada por el especialista en radiología. Ahora bien, indicaron que los valores absolutos de los especialistas en radiología en el estudio fueron menores a los observados en la práctica clínica en EE.UU. y Europa.

Salim *et al.* (35), en un estudio realizado también en Suecia (coincidente con la iniciativa de diálogo sobre evaluación y métodos de ingeniería inversa (DREAM por sus siglas en inglés) en DM (8.805 mujeres, 8,4 % cánceres), analizaron tres sistemas de IA (denominados IA 1, IA 2 e IA 3) comercialmente disponibles, aunque no identificados, obteniendo para el sistema IA 1 una sensibilidad superior a la del primer lector original

(81,9 % (78,9-84,6 %) vs 77,4 % (74,2-84,4 %), P = 0,03) y para los sistemas IA 2 e IA 3, una sensibilidad inferior a la del primer lector original (67,0 % (63,5-70,4 %) y 67,4 % (63,9-70,8 %) vs 77,4 %). Cuando la sensibilidad de los sistemas de IA se comparó con la de la lectura original del segundo lector (80,1 % (77,0-82,9 %)) los resultados se mantuvieron igual, aunque la diferencia observada frente al sistema IA 1 no fue significativa (P = 0,40). Todos los sistemas de IA tuvieron una sensibilidad inferior en comparación con la decisión original de consenso (85 % (82,2-87,5 %), siendo la diferencia observada no significativa (P = 0,11) cuando se comparó frente al sistema IA 1. En el estudio la especificidad de los algoritmos AI CAD se preseleccionó para que coincidiera con la especificidad del primer lector y, por tanto, no se comparó.

Por último, Schaffer *et al.* (36), tomando como referencia la inciativa DREAM en el que se analizaron 68.008 mujeres consecutivas procedentes del programa de cribado de Suecia, obtuvieron para el sistema de IA considerado el mejor modelo entre 31 sistemas de IA evaluados en la fase competitiva del reto DREAM sobre el conjunto de datos sueco, una especificidad inferior a la del primer lector original (88 % vs 96,7 %) y a la de la lectura original de consenso (81 % vs 98,5 %), para un umbral de predicción de la especificidad del algoritmo igual a la sensibilidad del primer lector original (77,1 %) y a la lectura original de consenso (83,9 %), respectivamente. Cuando analizaron el sistema de IA que denominaron CEM, consistente en una agregación ponderada de los ocho sistemas de IA con mejores resultados, obtuvieron que la especificidad de este sistema también fue inferior cuando se comparó con el primer lector original (92,5 % vs 96,7 %, P < 0,001) para un umbral de predicción de la especificidad del algoritmo igual a la sensibilidad del primer lector original (77,1 %).

IA como instrumento de ayuda a los especialistas en radiología

En los tres estudios (37-39) en los que se evaluaron sistemas de IA como instrumento de ayuda a los especialistas en radiología, la capacidad discriminativa de la lectura radiológica asistida con IA para la detección o no de cáncer de mama se consideró aceptable en dos estudios (38, 39) donde el valor de la AUC-ROC fue de 0,89 (0,85-0,92) y 0,8184, respectivamente, y no aceptable en uno (37), donde el valor AUC-ROC fue de 0,797 (0,754-0,840). Cuando se comparó con los valores AUC-ROC de la lectura radiológica no asistida con IA se observó que estos fueron inferiores a los anteriores, variación entre 0,769 (0,724-0,814) y 0,87 (0,83-0,90). En dos estudios (37, 38) se señaló que la diferencia encontrada fue estadísticamente significativa (P = 0,035 y P = 0,02, respectivamente).

En estos estudios la sensibilidad y la especificidad se comunicaron como una media de la lectura realizada con y sin la ayuda de la IA por 14 (37, 38) y de siete (39) especialistas en radiología. Los valores estimados indicaron que la sensibilidad media fue mayor para los especialistas en radiología con ayuda de la IA que para la lectura sin ayuda: 69,1 % (60,0-78,2 %) vs 65,8 % (57,4-74,3 %), P = 0,021 (37); 86 % (84-88 %) vs 83 % (81-85 %), P = 0,02 (38); y 62 % (41-75 %) vs 51 % (25-71 %), P = 0,03 (39). Con respecto a la especificidad la diferencia observada entre la lectura realizada por los especialistas en radiología con y sin ayuda no fue estadísticamente significativa en los dos estudios que la reportaron (P = 0,0634 (37) y P = 0,06 (38)). La generalización de estos resultados a la práctica clínica es limitada ya que en ellos se analiza la precisión de la lectura de los especialistas en radiología en un entorno de laboratorio mediante un conjunto de pruebas enriquecido.

IA como herramienta de clasificación previa al cribado

Como señalan Freeman *et al.* (5), cuando los sistemas de IA se utilizan como herramienta de clasificación previa al cribado, estos requieren una alta sensibilidad, de modo que se excluyan pocas pacientes con cáncer de la revisión radiológica, y una especificidad moderada, que determina la carga de casos radiológicos ahorrados. Cuatro estudios (40-43) evaluaron sistemas de IA comerciales como preselección para la identificación de mujeres de bajo riesgo cuyas mamografías requirieron poca o ninguna revisión radiológica.

Balta *et al.* (40) que evaluaron el sistema de IA Transpara v1.6.0 en una cohorte retrospectiva consecutiva alemana, obtuvieron una sensibilidad del 92,11 % y una especificidad del 65,50 % para un punto de corte ≤ 7, utilizado para eliminar pacientes de bajo riesgo de la doble lectura, y una sensibilidad del 95,61 % y una especificidad del 44,94 % para un punto de corte ≤ 5.

Por su parte, Dembrower *et al.* (41) evaluaron en una cohorte sueca un sistema de IA comercial (Lunit v5.5.0.16) como un sistema de preselección para eliminar pacientes considerados normales y posteriormente como un sistema postselección de pacientes negativos después de lectura doble para identificar cánceres de intervalo y cánceres detectados en la siguiente ronda de cribado. Utilizando un muestreo 11 veces superior de mujeres sanas para simular una población de cribado, informaron que el uso de la IA sin evaluación radiológica posterior en el 50 % y el 90 % de las mujeres con las puntuaciones de IA más bajas tuvo una sensibilidad del 100 % y el 96 % y una especificidad del 50 % y el 90 %, respectivamente.

Para el caso en el que el 2 % de las mujeres con las puntuaciones más altas de IA fueron postcribadas (con una hipotética prueba de seguimiento perfecta), la evaluación realizada por el sistema de IA de las mamografías negativas tras doble lectura detectó un 19 % de los 547 cánceres de intervalo y cánceres detectados en la siguiente ronda de cribado (32 cánceres de intervalo y 71 cánceres detectados en la siguiente ronda).

Lang *et al.* (42), para el sistema de IA Transpara v1.4.0 analizado en una cohorte sueca, alcanzaron una sensibilidad del 89,71 % y una especificidad del 53,35 % para un punto de corte ≤ 5. En ambos estudios, para un punto de corte ≤ 2 la sensibilidad fue del 100 % y la especificidad del 15 % y 19 %, respectivamente. En estos estudios la sensibilidad se refirió a la detección de cánceres realizada por los especialistas en radiología originales ya que las mujeres con cribado negativo no fueron objeto de seguimiento.

Por último, Raya-Povedano *et al.* (43), que analizaron el sistema Transpara v1.6.0 en una cohorte española, obtuvieron una sensibilidad del 88 % y una especificidad del 72 % para un punto de corte ≤ 7. Además, para la estrategia de clasificación con IA en comparación con el contexto de cribado de lectura doble, la sensibilidad fue no inferior (69,0 % (60,0-76,8 %) vs 67,3 % (58,2-75,2 %), P = 0,68), aunque sí la tasa de revaloración (4,2 % (3,9-4,5 %) vs 5,1 % (4,7-5,4 %), P < 0,001).

Ninguno de estos estudios aportó datos empíricos sobre el efecto en el comportamiento de los especialistas en radiología de la integración de la IA en la vía de cribado.

Carga de trabajo

En los estudios en los que los sistemas de IA se utilizaron como instrumento de ayuda al especialista en radiología, solo dos (37, 38) evaluaron la incidencia en la carga de trabajo. Para ello se midió el tiempo de lectura de estos con y sin ayuda de la IA.

En el primer estudio (37), el tiempo de lectura se midió para las dos sesiones evaluadas, separadas entre sí por un periodo de cuatro semanas, no incluyéndose las lecturas que fueron superiores a los 10 minutos. El resultado obtenido señaló un aumento en el tiempo de lectura medio cuando se utilizó la IA como ayuda al lector para ambas sesiones (sesión 1: 71,93 s vs 62,79 s, P < 0,001; sesión 2: 62,16 s vs 57,22 s, P < 0,001). Además, para la segunda sesión, analizaron el tiempo de lectura en función de las categorías de puntuación MammoScreen lo que mostró un efecto de aprendizaje. El efecto del aprendizaje ocasionó una disminución del tiempo de lectura

para una puntuación inferior a 4, y un aumento del tiempo de lectura de menos de 10 segundos para las puntuaciones superiores a 4.

En el segundo estudio (38), el tiempo de lectura se midió automáticamente por caso a través del software de la estación de trabajo. Al igual que en el estudio anterior, se observó un aumento en el tiempo de lectura cuando se realizó con ayuda de la IA (149 s vs 146 s, P = 0,15). También evaluaron la curva de aprendizaje del sistema de IA en relación con el tiempo de lectura. Observaron que la lectura sin ayuda y con apoyo de IA difería en función de la puntuación del ordenador Transpara (P < 0,001). Para exámenes de baja sospecha (puntuación, 1-5), los especialistas en radiología redujeron su tiempo medio de lectura por caso en un 11 % cuando utilizaron el sistema de IA, mientras que para los exámenes de alta sospecha (puntuación, 6-10) el tiempo de lectura por caso fue un 2 % mayor con el uso del soporte de IA.

Un estudio (43) en el que se evaluó la IA para clasificación analizó la carga de trabajo que ocasionó frente al contexto original (doble lectura). Para ello, la carga de trabajo del cribado se definió como el número de lecturas, y se calculó una estimación en horas utilizando el tiempo medio de lectura por examen comunicado originalmente en esta cohorte, 25 segundos para un examen de DM. Los resultados reflejaron que con el sistema de IA se redujo la carga de trabajo un 70,15 % (9.100 lecturas (63 horas) con IA vs 31.974 (222 horas), IC al 95 %: 70,6-72,4 %, P < 0,001).

Tabla 7. **Tabla de resultados de los estudios individuales incluidos en la RS de Freeman *et al.* (5)**

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
IA como sistema independiente							
Lotter *et al.* 2021 (32)	DM	*Cáncer índice* encontrados en la práctica del estudio del lector: mamografías de cribado adquiridas recientemente antes de la malignidad demostrada por biopsia.	**AUC:** AUC: 0,945 (0,919-0,968)	**Con la especificidad del especialista en radiología** *Sistema de AI in-house*: 96,2 % (91,7-99,2 %) *Comparador: lectura única*: 82,0 %	14,2 % (9,2-18,5 %), con la especificidad media del lector. < 0,0001	**Con la sensibilidad del especialista en radiología** *Sistema de AI in-house*: 90,9 % (84,9-96,1 %) *Comparador: lectura única*: 66,9 %	24,0 % 17,4-30,4 %, con la sensibilidad media del lector. < 0,0001
McKinney *et al.* 2020 (33)	DM	*Conjunto de datos R.U.* Comparación 1.er lector (valores obtenidos de la historia clínica) vs IA.	**AUC:** AUC: 0,889 (0,871-0,907)	*Sistema de AI in-house*: 65,42 % *Comparador 1.er lector*: 62,69 %	2,70 % –3,0-8,5 % 0,0043 No inferioridad	*Sistema de AI in-house*: 94,12 % *Comparador 1er lector*: 92,93 %	1,18 % 0,29-2,08 % 0,0096 Superioridad
		Conjunto de datos R.U. Comparación 2.º lector (valores obtenidos de la historia clínica) vs IA.		*Sistema de AI in-house*: 69,40 % *Comparador 2.º lector*: 69,40 %	0 % –4,89-4,89 % 0,0225 No inferioridad	*Sistema de AI in-house*: 92,13 % *Comparador 2º lector*: 92,97 %	–0,84 % –1,97-0,28 % No inferioridad
		Conjunto de datos R.U. Comparación consenso (valores obtenidos de la historia clínica) vs IA.		*Sistema de AI in-house*: 68,12 % *Comparador consenso*: 67,39 %	0,72 % –3,49-4,94 % 0,0039 No inferioridad	*Sistema de AI in-house*: 96,24 % *Comparador consenso*: 96,24 %	–3,35 % –4,06–-2,63 % No inferioridad

.../...

…/…

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
McKinney *et al.* 2020 (33) *(continuación)*	DM	*Conjunto de datos EE.UU.* Comparación consenso (valores obtenidos de la historia clínica) vs IA.	**AUC:** AUC: 0,8107 (0,791-0,831)	*Sistema de AI in-house*: 57,50 % *Comparador lectura única*: 48,10 %	9,40 % 4,45-13,85 % 0,0004 Superioridad	*Sistema de AI in-house*: 86,53 % *Comparador lectura única*: 80,83 %	5,70 % 2,62-6,4 % 0,0002 Superioridad
		Sistema de IA entrenado exclusivamente con datos R.U. y comprobado con el conjunto de datos EE.UU. (generalidad de la IA entre poblaciones y sistemas sanitarios).	Ne	*Sistema de AI in-house*: 56,24 % *Comparador lectura única*: 48,10 %	8,14 % 3,54-12,50 % 0,0006 Superioridad	*Sistema de AI in-house*: 84,29 % *Comparador lectura única*: 80,83 %	3,47 % 0,6-5,98 % 0,0212 Superioridad
		Comparación de la IA vs seis especialistas en radiología (MQSA) (interpretaron 500 mamografías elegidas aleatoriamente del grupo de pruebas EE.UU.).	**AUC:** *Sistema de AI:* AUC: 0,740 (0,696-0,794) *Media 6 especialistas en radiología:* AUC: 0,625 (0,032) Δ AUC: 0,115 (0,055-0,175) P = 0,0002	Ne	Ne	Ne	Ne

…/…

.../...

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Rodriguez-Ruiz *et al.* 2019 (34)	DM	Nueve conjuntos de datos. Sistema IA Transpara 1.4.0 comparado con 101 especialistas en radiología.	**AUC:** *Sistema IA Transpara*: AUC: 0,840 (0,820-0,860) *Especialistas en radiología:* AUC: 0,814 (0,787-0,841) *Diferencia IA Transpara vs especialistas en radiología:* AUC: 0,026 (–0,003-0,055)	*Sistema IA Transpara:* 75 % (65-85 %)-86 % (76-96 %) *Especialistas en radiología:* 76 % (67-85 %)-84 % (76-92 %)	*Sistema IA Transpara vs especialistas en radiología:* –2 % (–11,2-7 %)-8 % (–0,9-17,5 %)	*Para ambos:* 49 % (40-61 %)-79 % (73-86 %)	Ne
		Sistema IA Transpara 1.4.0 comparado con 101 especialistas en radiología. Base de datos C (Hupse *et al.*, 2013 Netherlands).	Ne	*Sistema IA Transpara:* 80 % (70-90 %) *Especialistas en radiología:* 77 % (70-83 %)	*Sistema IA Transpara vs especialistas en radiología:* 3 % (–6,2-12,6 %)	*Para ambos:* 77 % (70-83 %)	Ne
Salim *et al.* 2020 (35)	DM	*Conjunto de datos CSAW* (Swedish Cohort of Screen-Age Women). Tres sistemas de IA anónimos (IA 1, IA 2, IA 3).	**AUC:** *IA 1:* AUC: 0,956 (0,948-0,966) *IA 2:* AUC: 0,922 (0,910-0,934 *IA 3:* AUC: 0,920 (0,909-0,931) Diferencia IA 1 vs IA 2 e IA 3: P < 0,001 Diferencia IA 2 vs IA 3: P = 0,68	*IA 1:* 81,9 % (78,9-84,6 %) *IA 2:* 67,0 % (63,5-70,4 %) *IA 3:* 67,4 % (63,9-70,8 %) *1er lector:* 77,4 % (74,2-84,4 %) *2.º lector:* 80,1 % (77,0-82,9 %) *Lectura de consenso.* 85,0 % (82,2 %-87,5 %)	*IA 1 vs IA 2 e IA 3:* P < 0,001 *IA 1 vs 1.er lector:* P = 0,03 *IA 1 vs 2.º lector:* P = 0,40 *IA 1 vs lectura de consenso:* P = 0,11	*IA 1:* 96,6 % (96,5-96,7 %) *IA 2:* 96,6 % (96,5-96,7 %) *IA 3:* 96,7 % (96,6-96,8 %) *1.er lector:* 96,6 % (96,5-96,7 %) *2.º lector:* 97,2 % (97,1-97,3 %) *Lectura de consenso:* 98,5 % (98,4-98,6 %)	Ne
		Conjunto de datos CSAW. IA combinado.	Ne	*IA combinado:* 86,7 % (84,2-89,2 %)	*IA combinado vs IA 1:* P = 0,01	*IA combinado:* 92,5 % (92,3-92,7 %)	*IA combinado vs IA 1:* P < 0,001

.../...

.../...

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
Schaffter *et al.* 2020 (36)	DM	*Conjunto de datos KPW* (Kaiser Permanente Washington). Reto DM DREAM.	**AUC:** *Mejor modelo:* AUC: 0,855 *Mejor modelo mejorado:* 0,858	*Mejor modelo (sistema IA in-house):* 85,9 % (especialista en radiología) *Mejor modelo mejorado (sistema IA in-house):* 85,9 %	Ne	*Mejor modelo (sistema IA in-house):* 68,5 % *Mejor modelo mejorado (sistema IA in-house):* 66,3 %	Ne
		Conjunto de datos KI (Karolinska Institute). 20 mejores métodos.	**AUC:** *Mejor modelo:* AUC: 0,903	*Mejor modelo (sistema IA in-house):* 83,9 % (especialista en radiología)	Ne	*Mejor modelo (sistema IA in-house):* 81,2 %	Ne
		Conjunto de datos KPW. Modelo CEM.	**AUC:** *Modelo CEM:* AUC: 0,895	*Modelo CEM (sistema IA in-house):* 85,9 % (especialista en radiología)	Ne	*Mejor modelo (sistema IA in-house):* 66,3 % *CEM (sistema IA in-house):* 76,1 % *Especialista en radiología:* 90,5 %	Ne
		Conjunto de datos KI. Mejor modelo.	**AUC:** *Modelo CEM:* AUC: 0,903	*Mejor modelo (sistema IA in-house):* 77,1 % *Comparador: 1.er lector original:* 77,1 %	Ne	**Con la sensibilidad del 1.er lector original** *Mejor modelo (sistema IA in-house):* 88 % *Especialista en radiología:* 96,7 %	Ne

.../...

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
Schaffter *et al.* 2020 (36) *(continuación)*	DM	*Conjunto de datos KI*. Modelo CEM.	**AUC:** *Modelo CEM:* AUC: 0,923	*Modelo CEM (sistema IA in-house):* 77,1 % *Comparador: 1.er lector original:* 77,1 %	Ne	**Con la sensibilidad del 1.er lector original** *CEM (sistema IA in-house):* 92,5 % *Especialista en radiología:* 96,7 %	Ne
		Conjunto de datos KI. Mejor modelo (generalizable a doble lectura).	Ne	*Mejor modelo (sistema IA in-house):* 83,9 % *Comparador: lectura original de consenso*: 83,9 %	Ne	**Con la sensibilidad de la lectura original de consenso** *Mejor modelo (sistema IA in-house):* 81,2 % *Comparador: lectura original de consenso*: 98,5 %	Ne

.../...

.../...

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
IA como ayuda al lector							
Pacile *et al.* 2020 (37)	DM	Sistema IA MammoScreen versión 1. Comparador: lector único (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,797 (0,754-0,840) *Lectura radiológica no asistida con IA:* AUC: 0,769 (0,724-0,814) *Diferencia:* 0,028 (0,002-0,055) P = 0,035	*Lectura radiológica asistida con IA:* 69,1 % (60,0-78,2 %) *Lectura radiológica no asistida con IA:* 65,8 % (57,4-74,3 %)	3,3 % (1,7-7,2 %) P = 0,021	*Lectura radiológica asistida con IA:* 73,5 % (65,6-81,5 %) *Lectura radiológica no asistida con IA:* 72,5 % (65,6-79,4 %)	1,0 % (–3,0-3,8 %) P = 0,0634
			Tiempo de lectura: *Lectura radiológica asistida con IA (1.ª sesión):* 71,93 s (69,52-74,33) *Lectura radiológica no asistida con IA (1.ª sesión):* 62,79 s (60,77-64,80) *Diferencia:* P < 0,001 *Lectura radiológica asistida con IA (2ª sesión):* 62,16 s (60,04-64,29) *Lectura radiológica no asistida con IA (2ª sesión):* 57,22 s (55,10-59,33) *Diferencia:* P < 0,001				

.../...

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Rodriguez-Ruiz *et al.* 2019 (38)	DM	Sistema IA Transpara 1.3.0. Comparador lector único (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,89 (0,85-0,92) *Lectura radiológica no asistida con IA:* AUC: 0,87 (0,83-0,90) *Diferencia*: 0,02 (0,01-0,03) P = 0,02	*Lectura radiológica asistida con IA:* 86 % (84-88 %) *Lectura radiológica no asistida con IA:* 83 % (81-85 %)	3 % P = 0,02	*Lectura radiológica asistida con IA:* 79 % (77-81 %) *Lectura radiológica no asistida con IA:* 77 % (75-79 %)	2 % P = 0,06
			Tiempo de lectura: *Lectura radiológica asistida con IA:* 149 s (146-152) *Lectura radiológica no asistida con IA:* 146 s (143-147) *Diferencia:* P = 0,15				
Watanabe *et al.* 2019 (39)	DM	Sistema IA cmAssist. Comparador lector único (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,8148 *Lectura radiológica no asistida con IA:* AUC: 0,7599	*Lectura radiológica asistida con IA:* 62 % (41-75 %) *Lectura radiológica no asistida con IA:* 51 % (25-71 %)	11 % P = 0,03	Ne	Ne

.../...

…/…

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
IA para clasificación							
Balta *et al.* 2020 (40)	DM	Sistema IA Transpara 1.6.0.	Ne	*Puntuación riesgo ≤ 2 (aproximadamente el 15 % de bajo riesgo):* 100 %	Ne	*Puntuación riesgo ≤ 2 (aproximadamente el 19 % de bajo riesgo):* 15,49 %	Ne
			Ne	*Puntuación riesgo ≤ 5 (aproximadamente el 45 % de bajo riesgo):* 95,61 %	Ne	*Puntuación riesgo ≤ 5 (aproximadamente el 45 % de bajo riesgo):* 44,94 %	Ne
			Ne	*Puntuación riesgo ≤ 7 (aproximadamente el 65 % de bajo riesgo):* 92,11 %	Ne	*Puntuación riesgo ≤ 7 (aproximadamente el 65 % de bajo riesgo):* 65,50 %	Ne
Dembrower *et al.* 2020 (41)	DM	Sistema IA Lunit 5.5.0.16. Flujo de trabajo sin especialistas en radiología (predicción de cánceres detectados en el cribado).	**Número de cánceres detectados en el cribado que se perderían:** *Puntuación riesgo ≤ 0,0293 = puntuación de riesgo de la IA más bajo del 60 %:* 0	*Puntuación riesgo ≤ 0,0293 = puntuación de riesgo de la IA más bajo del 60 % (60 % de bajo riesgo):* 100 %	Ne	*Puntuación riesgo ≤ 0,0293 = puntuación de riesgo de la IA más bajo del 60 % (60 % de bajo riesgo):* 60,28 %	Ne
			Número de cánceres detectados en el cribado que se perderían: *Puntuación riesgo ≤ 0,0293 = puntuación de riesgo de la IA más bajo del 80 %:* 9	*Puntuación riesgo ≤ 0,0870 = puntuación de riesgo de la IA más bajo del 80 % (80 % de bajo riesgo):* 97,41 %	Ne	*Puntuación riesgo ≤ 0,0870 = puntuación de riesgo de la IA más bajo del 80 % (80 % de bajo riesgo):* 80,36 %	Ne
		Sistema IA Lunit 5.5.0.16. Flujo de evaluación mejorada (predicción de cánceres de intervalo).	**Detección potencial de cánceres de intervalo:** *Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 %:* 32	*Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 % (aproximadamente el 2 % de alto riesgo):* 16 %	Ne	*Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 % (aproximadamente el 2 % de alto riesgo):* 98,12 %	Ne

…/…

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
Dembrower *et al.* 2020 (41) *(continuación)*	DM	Sistema IA Lunit 5.5.0.16. Flujo de evaluación mejorada (predicción de cánceres de intervalo y de cánceres detectados en la siguiente ronda de cribado).	**Detección potencial de cánceres de intervalo y de cánceres detectados en la siguiente ronda de cribado:** *Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 %:* 103	*Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 % (aproximadamente el 2 % de alto riesgo):* 19 %	Ne	*Puntuación riesgo ≥ 0,5337 = puntuación de riesgo de la IA más alta del 2 % (aproximadamente el 2 % de alto riesgo):* 98,21 %	Ne
Lang *et al.* 2021(42)	DM	Sistema IA Transpara 1.4.0.	**Número de cánceres detectados en el cribado que se perderían:** *Puntuación riesgo ≤ 2*: 0	*Puntuación riesgo ≤ 2 (aproximadamente el 19 % de bajo riesgo):* 100 %	Ne	*Puntuación riesgo ≤ 2 (aproximadamente el 19 % de bajo riesgo):* 19,23 %	Ne
			Número de cánceres detectados en el cribado que se perderían: *Puntuación riesgo ≤ 5*: 7	*Puntuación riesgo ≤ 5 (aproximadamente el 53 % de bajo riesgo):* 89,71 %	Ne	*Puntuación riesgo ≤ 5 (aproximadamente el 53 % de bajo riesgo):* 53,35 %	Ne

.../...

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (*p*)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (*p*)
Raya-Povedano *et al.* 2021 (43)	DM	Sistema IA Transpara 1.6.0. Comparador. Doble lectura.	**Imágenes de bajo riesgo:** *Puntuación IA ≤ 7:* 71,5 %	*Sistema IA Transpara:* 69,0 % (60,0-76,8 %) *Doble lectura:* 67,3 % (58,2-75,2 %)	2,63 % (–4,9-11,4 %) P = 0,68 No inferioridad	Ne	Ne
			Tasa de revaloración: *Sistema IA Transpara*: 4,2 % (3,9-4,5 %) *Doble lectura*: 5,1 % (4,7-5,4 %) *Diferencia relativa*: –16,9 (–24,0-–11,0) P < 0,01				
			Carga de trabajo: *Sistema IA Transpara*: 9.100 lecturas (63 horas). *Doble lectura*: 31.974 lecturas (222 horas) *Diferencia relativa*: –71,5 (–72,4-–70,6) P < 0,01				

Tabla 8. **Tabla de contingencia 2x2 de los estudios individuales incluidos en la RS de Freeman *et al.* (5)**

Estudio	Prueba índice (proveedor) / comparador	Casos	Cáncer	IA (*stand-alone*) (DM)					
				Sensibilidad	Especificidad	TP	FP	FN	TN
IA como sistema autónomo									
Lotter *et al.* (32)	IA *in-house*	285	131	0,962	0,669	126	51	5	103
	IA *in-house*			0,82	0,909	107	14	24	140
	Comparador. Lector único (media)			0,82	0,669	107	51	24	103
McKinney *et al.* (33)	IA *in-house*	3.097	686	0,562	0,843	386	379	300	2.032
	Comparador. Lector único original			0,481	0,808	330	462	356	1.949
Rodriguez-Ruiz *et al.* (34)	IA (Transpara v1,4,0)	199	79	0,8	0,79	63	25	16	95
	Comparador. Lector único (media)			0,77	0,79	61	25	18	95
Salim *et al.* (35)	IA-1 (anónimo)	113.663	739	0,819	0,966	605	3.839	134	109.085
	IA-2 (anónimo)			0,67	0,966	495	3.839	244	109.085
	IA-3 (anónimo)			0,674	0,967	498	3.726	241	109.198
	Comparador. Primer lector original			0,774	0,966	572	3.839	167	109.085
	Comparador. Segundo lector original			0,801	0,972	592	3.162	147	109.762
	Comparador. Lectura original de consenso			0,85	0,985	628	1.694	111	111.230

…/…

…/…

Estudio	Prueba índice (proveedor) / comparador	Casos	Cáncer	IA (*stand-alone*) (DM)					
				Sensibilidad	Especificidad	TP	FP	FN	TN
Schaffer *et al.* (36)	IA *in-house*. Mejor modelo	68.008	780	0,771	0,88	601	8.067	179	59.161
	IA *in-house*. Modelo CEM			0,771	0,925	601	5.042	179	62.186
	Comparador. Primer lector original			0,771	0,967	601	2.219	179	65.009
	Comparador. Lectura original de consenso			0,839	0,985	654	1.008	126	66.220
IA como ayuda al lector									
Pacile *et al.* (37)	IA (MammoScreen v1)	240	120	0,691	0,735	83	32	37	88
	Comparador. Lector único (media)			0,658	0,725	79	33	41	87
Rodriguez-Ruiz *et al.* (38)	IA (Transpara v1.3.0)	240	100	0,86	0,79	86	29	14	111
	Comparador. Lector único (media)			0,83	0,77	83	32	17	108
Watanabe *et al.* (39)	IA (cmAssist)	122	90	0,62	0,772	56	7	34	25
	Comparador. Lector único (media)			0,51	0,781	46	7	44	25
IA para clasificación									
Balta *et al.* (40)	IA (Transpara v1.6.0)								
	Puntuación de la IA ≤ 2: ~ 15 % bajo riesgo	17.890	114	1	0,155	114	15.028	0	2.754
	Puntuación de la IA ≤ 5: ~ 45 % bajo riesgo			0,956	0,449	109	9.791	5	7.991
	Puntuación de la IA ≤ 7: ~ 65 % bajo riesgo			0,921	0,655	105	6.135	9	11.647

…/…

.../...

Estudio	Prueba índice (proveedor) / comparador	Casos	Cáncer	IA (*stand-alone*) (DM)					
				Sensibilidad	Especificidad	TP	FP	FN	TN
Dembrower *et al.* (41)	IA (Lunit v5.5.0.1.6)								
	Puntuación IA ≥ 0,0293: 60 % de bajo riesgo	75.334	347	1	0,603	347	29.787	0	45.200
	Puntuación IA ≤ 0,0870: 80 % de bajo riesgo	75.334	347	0,974	0,804	338	14.729	9	60.258
Lang *et al.* (42)	IA (Transpara v1.4.0)								
	Puntuación de la IA ≤ 2: ~ 19 % bajo riesgo	9.581	68	1	0,192	68	7.684	0	1.829
	Puntuación de la IA ≤ 5: ~ 53 % bajo riesgo			0,897	0,533	61	4.438	7	5.075
	Puntuación de la IA ≤ 7: ~ 73 % bajo riesgo			0,838	0,733	57	2.541	11	6.972
Raya-Povedano *et al.* (43)	IA (Transpara v1.6.0)								
	Puntuación de la IA ≤ 7: ~ 72 % bajo riesgo	15.987	113	0,885	0,720	100	4.437	13	11.437

II.3.1.4.2. Descripción y análisis de los resultados de los estudios individuales

Los resultados de los estudios individuales analizados en este informe quedan descritos en las tablas 9 y 10. En ambas tablas los resultados se muestran para los distintos papeles potenciales que la IA puede jugar en el proceso de cribado de cáncer de mama: como lector autónomo en sustitución de alguno de los especialistas en radiología, como instrumento de apoyo a los especialistas en radiología y como herramienta de clasificación previa al cribado.

IA como sistema autónomo para sustituir alguno de los especialistas en radiología

Tres estudios (26-28) en los que se evaluó la IA como instrumento de apoyo a los especialistas en radiología, proporcionaron el valor del parámetro AUC-ROC. En dos (27, 28) la capacidad discriminativa de los sistemas de IA se consideró buena, AUC-ROC igual a 0,951 (0,947-0,955) (27) e igual a 0,93 (0,89-0,96) (28), y en uno (26) aceptable, AUC-ROC igual a 0,852 (0,836-0,869). Además, un estudio (28) también obtuvo el valor AUC-ROC para DBT, el cual mostró una capacidad discriminativa buena, igual a 0,94 (0,91-0,97).

Hsu *et al.* (26) evaluaron un modelo *in-house* de IA denominado CEM (ya analizado en el estudio de Schaffer *et al.* de 2020 (36)) surgido de la fase competitiva DREAM en el conjunto de datos UCLA obtenido de cinco centros médicos de la Universidad de California. Dicho conjunto de datos fue sobremuestreado para exámenes considerados FP, VP y FN y submuestreado para VN. Como resultado obtuvieron que tanto la sensibilidad como la especificidad del modelo CEM fue inferior comparada con los resultados de los especialistas en radiología. Para una especificidad igual a la de la lectura única realizada por especialistas en radiología de 93,0 % (92,9-93,2 %), la sensibilidad fue del 54,7 % (50,8-58,8 %) para el sistema de IA y del 82,6 % (79,5-85,6 %) para la lectura radiológica (diferencia –27,8 % (–32,2-–23,3 %, P < 0,001), y para una sensibilidad igual a la de los especialistas en radiología, 82,6 % (79,5-85,6 %), la especificidad fue del 69,7 % (63,7-74,9 %) para la IA y del 93,0 % (92,9-93,2 %) para la lectura única radiológica (diferencia –23,4 % (–29,5-–18,2 %, P < 0,001).

Leibig *et al.* (27) alcanzaron unos resultados similares a los anteriores. Para un sistema de IA *in-house* evaluado en un conjunto de datos enriquecidos obtenidos de dos centros de cribado en Alemania obtuvieron una sensibilidad del 84,6 % (83,3-85,9 %) y una especificidad del 91,3 %

(91,1-91,5 %) menor que la sensibilidad (87,2 % (85,6-88,7 %)) y especificidad (93,4 % (93,2-93,6 %)) de la lectura original del especialista en radiología. Las diferencias observadas fueron estadísticamente significativas, P = 0,0019 y P < 0,0001.

Sharma *et al.* (29) para muestras no enriquecidas representativas de 10 y un año, creadas a partir de una muestra histórica consecutiva de casos procedentes de tres centros del R.U. y de uno de Hungría, señalaron para la muestra de 10 años una sensibilidad del sistema de IA superior a la de la lectura histórica del primer lector, 78,1 % (76,6-79,7 %) vs 76,4 % (74,9-78,0 %), diferencia 1,7 % (0,1-3,3 %) y una especificidad inferior, 91,2 % (91,0-91,4 %) vs 96,0 % (95,9-96,2 %), diferencia –4,8 % (–5,1-–4,6 %). Para la muestra de un año, en la que se dispuso de datos de cánceres de intervalo más completos, también se observó una mayor sensibilidad (diferencia 5,1 %) y una menor especificidad (diferencia –5,2 %) para el sistema de IA en comparación con la lectura del primer especialista en radiología. Que la especificidad para el sistema de IA fuese menor contribuyó a una mayor necesidad de consenso para la doble lectura. Además, señalaron que en la muestra de 10 años el sistema de IA detectó el 29,8 % (111 de 373) de los cánceres de intervalo históricos y en la muestra de un año el 35,9 % (46 de 128).

Por último, Romero-Martín *et al.* (28), para una cohorte de cribado real procedente del ensayo Tomosynthesis Cordoba Screening Trial, obtuvieron para un punto de corte de detección del sistema de IA igual a 80 una sensibilidad del 62,8 % (53,6-71,2 %) para la IA y del 58,4 % (49,2-67,1 %) para la lectura única radiológica, diferencia igual a 4,4 % (–4,4-13,3 %), P = 0,458, no inferior. Para un punto de corte de 74, la diferencia observada entre la sensibilidad del sistema de IA y de la lectura doble fue igual a 3,5 % (–4,4-11,5 %), P = 0,523, no inferior. También percibieron que la tasa de revaloraciones disminuyó con el sistema de IA cuando se comparó con lectura única (punto de corte de 80) y lectura doble (punto de corte de 74) en –1,4 % (271 de 15.999 vs 498 de 15.999, P < 0,001) y en –2 % (493 de 15.999 vs 808 de 15.999, P < 0,001), respectivamente. Además, para exámenes de cribado con DBT, la sensibilidad del sistema de IA fue no inferior a la de la lectura única para un punto de corte de 65, ni a la lectura doble para un punto de corte de 57, diferencia para ambos igual a 3,5 % (–3,5-10,6 %), P = 0,648 y P = 0,481, respectivamente. La tasa de revaloraciones para exámenes DBT fue superior para el sistema de IA en comparación con la lectura única individual (6,2 % (5,7-6,7 %), P < 0,001) y con la lectura doble lectura original humana (12,3 % (11,7-12,9 %, P < 0,001), para los puntos de corte señalados.

IA como instrumento de ayuda a los especialistas en radiología

La capacidad discriminativa de los sistemas de IA analizados para DM como ayuda a la lectura radiológica se consideró buena en un estudio (21) (AUC-ROC igual a 0,91 (0,88-0-93), adecuada en dos (23,24)) (AUC-ROC igual a 0,858 (0,809-0,907) y a 0,852 (0,853-0,868), respectivamente) y no adecuada en uno (22) (AUC-ROC igual a 0,773 (0,723-0,823). Cuando se comparó la capacidad discriminativa de la lectura radiológica con ayuda de la IA frente a lectura única radiológica sin ayuda de la IA, esta fue superior para la primera, siendo la diferencia encontrada estadísticamente significativa en todos los estudios ($P < 0,01$ (21), $P = 0,004$ (22), $P < 0,001$ (23) y $P = 0,005$ (24). Para los análisis realizados con DBT, en un estudio (25) se observó una capacidad discriminativa adecuada, aumentando la AUC-ROC media de 0,833 (0,799-0,867) a 0,863 (0,829-0,898), $P = 0,0025$.

Cuatro estudios (21-24) analizaron para DM la especificidad y sensibilidad de la lectura radiológica con y sin ayuda de los sistemas de IA. En todos ellos se observó que la sensibilidad de la lectura radiológica única con ayuda de los sistemas de IA evaluados fue superior a la lectura única sin ayuda de IA, 95,07 % (92,08-97,01 %) vs 84,77 % (80,46-88,29 %) (21), 70 % (68-72 %) vs 66 % (63-70 %) (22), 83,0 % (80,7-85,3 %) vs 62,9 % (59,9-65,9 %) (23) y 68,78 % ± 18,67 vs 68,70 % ± 16,34 (24), siendo la diferencia observada entre ellas estadísticamente significativa en dos estudios, $P < 0,01$ (21) y $P < 0,001$ (23) y no significativa en otros dos, $P = 0,051$ (22) y $P = 0,937$ (24). Para estos mismos estudios, la especificidad varió de unos a otros. En dos fue mayor para la lectura con ayuda de IA frente a lectura sin ayuda de IA, 81 % (77-86 %) vs 79 % (74-85 %) (22) y 88,34 % ± 6,93 vs 82,05 % ± 4,65 (24), y en dos menor, 54,88 % (47,97-61,62 %) vs 63,55 % (56,68-69,93 %) (21) y 67,2 % (64,3-70,1 %) vs 68,7 % (65,8-71,6 %) (23). Las diferencias observadas con respecto a la especificidad solo fueron estadísticamente significativas para el estudio de Sun *et al.* (24), $P = 0,005$.

Un estudio realizado por vanWinkel *et al.* (25) analizó para DBT la especificidad y sensibilidad de la lectura radiológica asistida por sistemas de IA frente a la lectura no asistida por IA. En él, se obtuvo que la sensibilidad de la lectura con ayuda de IA fue superior en comparación con la lectura sin ayuda de IA, 79,2 % (73,3-85,1 %) vs 74,6 % (68,3-80,8 %), $P = 0,016$, mientras que la especificidad fue no superior, $P = 0,380$.

La generalización de estos resultados a la práctica clínica se consideró que fue limitada ya que en estos estudios se analizó la precisión de la lectura de los especialistas en radiología en un entorno de laboratorio mediante un conjunto de pruebas enriquecido.

IA como herramienta de clasificación previa al cribado

Como ya se ha señalado con anterioridad, el empleo de la IA como herramienta de clasificación previa al cribado requiere que la sensibilidad de los sistemas de IA sea alta, de modo que se excluyan pocas pacientes con cáncer de la revisión radiológica, y la especificidad moderada, que determina la carga de casos radiológicos ahorrados.

Tres estudios (27, 30, 31) evaluaron sistemas de IA comerciales (Transpara v1.7.0) o *in-house* como preselección para la identificación de mujeres de bajo riesgo cuyas mamografías requirieran poca o ninguna revisión radiológica. En el estudio de Larsen *et al.* (30) en el que se evaluó una cohorte retrospectiva de mujeres participantes en el programa poblacional de cribado de mama noruego con el sistema de IA Transpara v1.7.0., se exploró el funcionamiento de un sistema de IA como una herramienta de decisión binaria con tres diferentes umbrales, definidos prospectivamente, para la selección de imágenes sospechosas o no sospechosas de malignidad. Con el umbral 1, una puntuación bruta > 9 (puntuación del sistema de IA de 10) fue definida como «seleccionado» por el sistema de IA y una puntuación < 10 como «no seleccionado»; el umbral 2, igual a una puntuación bruta > 9,13, representó una tasa de selección igual a la de consenso (8,8 %) y se utilizó para explorar el funcionamiento de la IA cuando el número de exámenes seleccionados por el sistema como sospechosos fue similar al número de exámenes seleccionados por los dos especialistas en radiología; y el umbral 3, igual a una puntuación bruta > 9,43, correspondió con una tasa de selección igual a la tasa individual media de interpretaciones positivas observadas por los especialistas en radiología (del 5,8 %) en la muestra del estudio. Para el umbral 1 la sensibilidad fue del 77,8 % y la especificidad del 90,5 %, para el umbral 2 fue del 75,8 % y del 94,8 % y para el umbral 3 del 69,5 % y del 94,7 %. Con el umbral 1 se seleccionaron el 86,8 % de los cánceres detectados por el cribado (653 de 752) y el 93 % de los cánceres de intervalo invasivos (86 de 92); con el umbral 2 el 85,1 % de los cánceres detectados por el cribado (640 de 752) y el 41,5 % de los cánceres de intervalo (85 de 205); y con el umbral 3 el 80,1 % de los cánceres detectados por el cribado y el 30,7 % de los cánceres de intervalo (63 de 205). Además, con los umbrales 2 y 3, el 42,9 % (48 de 112) y el 43,3 % (65 de 150) de los cánceres no detectados mediante el cribado por el sistema de IA tuvieron una interpretación positiva por uno de los dos especialistas en radiología.

Lauritzen *et al.* (31) evaluaron en una cohorte consecutiva de mujeres cribadas bienalmente de cáncer de mama en la Región Capital de Dinamarca un sistema de IA (Transpara v1.7.0) para categorizar las mamografías como normales, de riesgo moderado o sospechosas. Previamente al análisis definieron un umbral de exclusión de 5, que significaba que aproxi-

madamente el 50 % de las mamografías se podrían considerar como normales. Además, utilizó un UR del 9,989 para determinar cuándo una mamografía se categorizaba como sospechosa. Así, para una puntuación de corte < 5 la mamografía se consideró normal y fue excluida de la lectura por los especialistas en radiología; para una puntuación de corte ≥ 5 o un UR ≤ 9,989 se consideró de riesgo moderado y fue leída por dos especialistas en radiología (las decisiones de revaloración de los especialistas en radiología se extrajeron de los informes de cribado originales); y para un UR > 9,989 se consideró sospechosa y fue excluida de la lectura por los especialistas en radiología. Los resultados del estudio señalaron que, de las 114.421 mamografías analizadas, 71.499 exámenes de cribado se etiquetaron como normales, 42.836 de riesgo moderado y 86 como sospechosas, siendo estas últimas revaloradas automáticamente. Además, simulando el cribado con IA en la muestra de ensayo, obtuvieron una sensibilidad del 69,7 % (66,9-72,4 %) y una especificidad el 98,6 % (98,5-98,7 %) para la IA frente a una sensibilidad del 70,8 % (68,0-73,5 %) y una especificidad del 98,1 % (98,1-98,2 %) para la lectura radiológica, diferencia en sensibilidad y especificidad, P = 0,02 (no inferior) y P < 0,001, respectivamente.

Leibig *et al.* (27) evaluaron lo que denominaron «vía de derivación de decisiones» en la que se utilizó un sistema de IA para clasificar un estudio como normal o sospechoso y para proporcionar al mismo tiempo una indicación de la confianza en su clasificación. Se determinaron unos umbrales para permitir categorizar los estudios que pasaban por el proceso de derivación de decisiones, en clasificación normal, red de seguridad y derivación al especialista en radiología. Los umbrales se representaron como conjuntos de dos puntos operativos, la sensibilidad del algoritmo en el conjunto de datos de validación más la especificidad del algoritmo en el conjunto de datos de validación. Para un umbral igual a una sensibilidad y especificidad del algoritmo en el conjunto de datos de validación del 97 % y del 98 %, respectivamente (NT@97 %+SN@98 %), la sensibilidad ((89,8 % (88,5-91,1 %)) y la especificidad (94,3 % (94,2-94,5 %)) del sistema de IA fue mayor que la sensibilidad (87,2 % (85,6-88,7 %)) y especificidad (93,4 % (93,2-93,6 %)) de la lectura independiente realizada por dos especialistas en radiología, diferencia en sensibilidad 2,6 %, P < 0,0001, y en especificidad 1,0 %, P < 0,0001. Para este umbral considerado el punto de funcionamiento modelo, el número de estudios correctamente etiquetados como normal fue del 63 %.

Carga de trabajo

La carga de trabajo se midió principalmente en los estudios en los que la IA se empleó como sistema de ayuda al especialista en radiología. Bao *et al.* (21) determinaron automáticamente mediante una estación de trabajo virtual el tiempo de lectura con y sin ayuda de la IA, el cual se de-

finió como el intervalo de tiempo entre el momento en el que se enviaron correctamente los registros y se completó la carga de las imágenes. Los resultados de este estudio indicaron una disminución en el tiempo de lectura cuando este se realizó con ayuda de la IA, 106 s por mamografía, frente a los 215 s empleados para la lectura sin ayuda, diferencia –109 s, $P < 0,01$. Dicha diferencia se volvió a observar cuando se midió el tiempo de lectura para los casos de cáncer (101 s por mamografía vs 221 s) y para los casos normales (112 s por mamografía vs 202 s).

Sun *et al.* (24) también observaron que el tiempo de lectura medio, automáticamente registrado para cada caso por el sistema, fue significativamente menor ($P = 0,032$) con ayuda de la IA (62,28 s ± 23,12 s) que sin ayuda (80,18 s ± 33,26 s). Además, analizaron la relación entre el tiempo de lectura y la puntuación de dificultad de las imágenes que ofrecía la IA. Para una puntuación de 1 a 5, dificultad baja, el tiempo de lectura media se redujo un 35,2 % y para un coeficiente de dificultad alta, de 6 a 9, el tiempo medio de lectura de cada caso aumentó en un 6,5 %. Dang *et al.* (22), que también midieron el tiempo automáticamente, no encontraron diferencia en el tiempo de lectura empleado por los especialistas en radiología con ayuda de la IA y sin ayuda, 101,8 s (80,8 s - 122,7 s) frente a 106,4 s (82,3 s - 130,5 s), diferencia –4,6 s, $P = 0,754$. Este resultado se mantuvo cuando la lectura la realizaron especialistas en radiología seniors y juniors. Por último, en el estudio de Lee *et al.* (23) en el que el tiempo de lectura se registró automáticamente mediante un sistema de informes basado en la web desde la visualización inicial de la imagen por parte del lector hasta la decisión final, el tiempo medio de lectura con ayuda de IA fue menor que sin ayuda (73,04 s vs 82,73 s, $P < 0,001$) cuando la lectura la realizaron especialistas en radiología de mama, y mayor (42,52 s vs 35,44 s, $P < 0,001$) cuando la lectura se realizó por especialistas en radiología generales.

VanWinkel *et al.* (25) evaluaron la carga de trabajo en mamografías DBT. Señalaron que el tiempo de lectura para exámenes DBT con ayuda de la IA fue menor que sin ayuda (36 s (35-37 s) vs (41 s (39-42 s), diferencia –5 s, $P < 0,001$). Esta diferencia se mantuvo independientemente de la densidad mamaria y del protocolo de lectura. La reducción en el tiempo de lectura con IA fue aún mayor cuando se utilizaron imágenes 2D de mamografía sintética y navegación interactiva que cuando no se utilizaron.

Por último, Lauritzen *et al.* (31) analizaron la carga de trabajo en un estudio en el que se empleó un sistema de IA para clasificación. En este, se consideró la reducción de la carga de trabajo como el porcentaje de mamografías leídas únicamente por el sistema de IA, correspondientes a mamografías normales o sospechosas (los especialistas en radiología sólo leerían las mamografías de riesgo moderado). El resultado indicó una disminución de la carga de trabajo del 62,6 % (lecturas evitadas: 71.585 de 114.421).

Tabla 9. **Tabla de resultados de los estudios individuales incluidos en el informe**

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
IA como sistema independiente							
Hsu *et al.* 2022 (26)	DM	Sistema IA *in house*, modelo CEM. Comparador: lectura especialista en radiología única.	**AUC:** IA AUC: 0,852 (0,836-0,869)	**Con la especificidad del especialista en radiología** *Sistema IA propio:* 54,7 % (50,8-58,8 %) *Lectura única:* 82,6 % (79,5-85,6 %)	–27,8 % (–32,2–-23,3 %) P < 0,001	**Con la sensibilidad del especialista en radiología** *Sistema IA propio*: 69,7 % (63,7-74,9 %) *Lectura única:* 93,0 % (92,9-93,2 %)	–23,4 % (–29,5–-18,2 %) P < 0,001
Leibig *et al.* 2022 (27)	DM	Sistema IA *in-house* comparado con lectura independiente realiza por 2 especialistas en radiología.	**AUC:** IA AUC: 0,951 (0,947-0,955)	*Sistema IA propio:* 84,6 % (83,3-85,9 %) *2 especialistas en radiología lectura independiente*: 87,2 % (85,6-88,7 %)	–2,6 % P = 0,0019	*Sistema IA propio:* 91,3 % (91,1-91,5 %) *2 especialistas en radiología lectura independiente:* 93,4 % (93,2-93,6 %)	–2,0 % P < 0,0001

…/…

…/…

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Romero-Martin *et al.* 2022 (28)	DM	Sistema IA Transpara 1.7.0. Comparador. Lectura única. *Puntuación de corte 80.*	**AUC:** IA AUC: 0,93 (0,89-0,96)	*Sistema IA Transpara*: 62,8 % (53,6-71,2 %) *Especialista en radiología lectura única:* 58,4 % (49,2-67,1 %)	4,4 % (–4,4-13,3 %) P = 0,458 No inferior	Ne	Ne
		Sistema IA Transpara 1.7.0. Comparador. Doble lectura. *Puntuación de corte 74.*		*Sistema IA Transpara:* 70,8 % (61,8-78,4 %) *Especialista en radiología doble lectura:* 67,3 % (58,2-75,2 %)	3,5 % (–4,4-11,5 %) P = 0,523 No inferior	Ne	Ne
	DBT	Sistema IA Transpara 1.7.0 Comparador. Lectura única. *Puntuación de corte 65.*	**AUC:** IA AUC: 0,94 (0,91-0,97)	*Sistema IA Transpara:* 80,5 % (72,3-86,8 %) *Especialista en radiología lectura única:* 77,0 % (68,4-83,8 %)	3,5 % (–3,5-10,6 %) P = 0,648 No inferior	Ne	Ne
		Sistema IA Transpara 1.7.0 Comparador. Doble lectura. *Puntuación de corte 57.*		*Sistema IA Transpara:* 85,0 % (77,2-90,4 %) *Especialista en radiología doble lectura:* 81,4 % (73,3-87,5 %)	3,5 % (–3,5-10,6 %) P = 0,481 No inferior	Ne	Ne

…/…

…/…

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Sharma *et al.* 2021 (29)	DM	*Muestra representativa de «10 años».* Sistema independiente IA MIA 2.0.1 Comparador. Primer lector.	NE	*Sistema IA MIA:* 78,1 % (76,6-79,7 %) *Primer lector:* 76,4 % (74,9-78,0 %)	1,7 % (0,1-3,3 %)	*Sistema IA MIA:* 91,2 % (91,0-91,4 %) *Primer lector:* 96,0 % (95,9-96,2 %)	–4,8 % (–5,1-–4,6 %)
		Muestra representativa de «1 año». Sistema independiente IA MIA 2.0.1. Comparador. Primer lector.	NE	*Sistema IA MIA:* 75,2 % (71,3-79,0 %) *Primer lector:* 70,1 % (66,1-74,1 %)	5,1 %	*Sistema IA MIA:* 91,4 % (91,0-91,9 %) *Primer lector:* 96,6 % (96,3-97,0 %)	–5,2 %
IA como ayuda al lector							
Bao *et al.* 2022 (21)	DM	Sistema IA Yizhun versión 3.2.3. Comparador. Lectura única de cualquiera de los dos especialistas en radiología de forma independiente.	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,91 (0,88-0,93) *Lectura radiológica no asistida con IA*: AUC: 0,84 (0,81-0,87) *Diferencia*: 0,07) P < 0,01	*Lectura radiológica asistida con IA:* 95,07 % (92,08-97,01 %) *Lectura radiológica no asistida con IA:* 84,77 % (80,46-88,29 %)	10,3 % P < 0,01	*Lectura radiológica asistida con IA:* 54,88 % (47,97-61,62 %) *Lectura radiológica no asistida con IA:* 63,55 % (56,68-69,93 %)	–8,67 % P = 0,07
			Carga de trabajo: *Lectura radiológica asistida con IA:* Tiempo de lectura: 106 s *Lectura radiológica no asistida con IA:* Tiempo de lectura: 215 s Diferencia: –109 s P < 0,01				

…/…

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Dang *et al.* 2022 (22)	DM	Sistema IA Mammoscreen v.1.2 Comparado. Lectura única (media). Análisis 3-CAT.BI-RADS. Umbral BI-RADS ≥ 3.	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,773 (0,723-0,823) *Lectura radiológica no asistida con IA:* AUC: 0,739 (0,689-0,789) *Diferencia*: 0,034 (0,012-0,056) P = 0,004	*Lectura radiológica asistida con IA:* 70 % (68-72 %) *Lectura radiológica no asistida con IA:* 66 % (63-70 %)	4 % (–2-8 %) P = 0,051	*Lectura radiológica asistida con IA:* 81 % (77-86 %) *Lectura radiológica no asistida con IA:* 79 % (74-85 %)	2 % (–5-9 %) P = 0,570
			Carga de trabajo: *Lectura radiológica asistida con IA:* Tiempo de lectura: 101.810 s (80.850-122.760 s) *Lectura radiológica no asistida con IA:* Tiempo de lectura: 106.410 s (82.320-130.520 s) *Diferencia*: –4.620 s P = 0,754				

.../...

.../...

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Lee *et al.* 2022 (23)	DM	Sistema IA Lunit INSIGHT MMG v 1.1.1.0 Comparador. Lectura única (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,858 (0,809-0,907) *Lectura radiológica no asistida con IA:* AUC: 0,748 (0,686-0,811) *Diferencia*: P < 0,001	*Lectura radiológica asistida con IA:* 83,0 % (80,7-85,3 %) *Lectura radiológica no asistida con IA:* 62,9 % (59,9-65,9 %)	P < 0,001	*Lectura radiológica asistida con IA:* 67,2 % (64,3-70,1 %) *Lectura radiológica no asistida con IA:* 68,7 % (65,8-71,6 %)	P = 0,273
			Carga de trabajo: <u>*Lectura realizada por especialistas en radiología especialistas en mama*</u> *Lectura radiológica asistida con IA:* Tiempo medio de lectura: 73,04 s. *Lectura radiológica no asistida con IA* Tiempo medio de lectura: 82,73 s. P < 0,001 <u>*Lectura realizada por especialistas en radiología generales*</u> *Lectura radiológica asistida con IA:* Tiempo medio de lectura: 42,52 s. *Lectura radiológica no asistida con IA* Tiempo medio de lectura: 35,44 s. P < 0,001				

.../...

…/…

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
Sun *et al.* 2021 (24)	DM	Sistema IA *in-house* Comparador. Lectura única (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,852 (0,835-0,868) *Lectura radiológica no asistida con IA:* AUC: 0,805 (0,786-0,823) *Diferencia*: P = 0,005	*Lectura radiológica asistida con IA:* 68,78 % ±18,67 *Lectura radiológica no asistida con IA:* 68,70 % ±16,34	P = 0,937	*Lectura radiológica asistida con IA:* 88,34 % ±6,93 *Lectura radiológica no asistida con IA:* 82,05 % ±4,65	P = 0,005
			Carga de trabajo: *Lectura radiológica asistida con IA:* Tiempo medio de lectura: 62,28 s ±23,12 *Lectura radiológica no asistida con IA:* Tiempo medio de lectura: 80,18 s ±33,26 P = 0,032				

…/…

…/…

Estudio	Modalidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
vanWinkel *et al.* 2022 (25)	BDT	Sistema IA Transpara 1.6.0. Comparador. Lectura úncia (media).	**AUC:** *Lectura radiológica asistida con IA:* AUC: 0,863 (0,829-0,898) *Lectura radiológica no asistida con IA:* AUC: 0,833 (0,799-0,867) *Diferencia*: 0,030 (0,011-0,049) P = 0,0025	*Lectura radiológica asistida con IA:* 79,2 % (73,3-85,1 %) *Lectura radiológica no asistida con IA:* 74,6 % (68,3-80,8 %)	6,2 % (1,3-11,1 %) P = 0,016	NE	1,1 % (–1,3-3,5 %) P = 0,380
			Carga de trabajo: *Lectura radiológica asistida con IA:* Tiempo de lectura: 36 s (35-37 seg) *Lectura radiológica no asistida con IA:* Tiempo de lectura: 41 s (39-42 seg) *Diferencia*: –5 s P < 0,001				

…/…

.../...

Estudio	Moda-lidad			Sensibilidad (95 % IC)	Δ Sensibilidad (95 % IC) (p)	Especificidad (95 % IC)	Δ Especificidad (95 % IC) (p)
IA para clasificación							
Larsen *et al.* 2022 (30)	DM	Sistema IA Transpara 1.7.0	Ne	*Puntuación de la IA = 10* 77,8 %	Ne	*Puntuación de la IA = 10* 90,5 %	Ne
			Ne	*Puntuación de la IA > 9,13* 75,8 %	Ne	*Puntuación de la IA > 9,13* 91,8 %	Ne
			Ne	*Puntuación de la IA > 9,43* 69,5 %	Ne	*Puntuación de la IA > 9,43* 94,7 %	Ne
Lauritzen *et al.* 2022 (31)	DM	Sistema IA Transpara 1.7.0 comparado con lectura independiente por dos especialistas en radiología. Puntuación de corte: < 5, normal; ≥ 5 o ≤ 9.989 (umbral revaloración), riesgo moderado; > 9,989, sospechosa.	**AUC:** IA AUC. 0,78 (0,77-0,79)	*Sistema IA:* 69,7 % (66,9-72,4 %) *Dos especialistas en radiología lectura independiente:* 70,8 % (68,0-73,5 %)	P = 0,02 No inferior	*Sistema IA:* 98,6 % (98,5-98,7 %) *2 especialistas en radiología lectura independiente:* 98,1 % (98,1-98,2 %)	P < 0,0001
			Carga de trabajo: –62,6 % (lectura evitada 71.585 de 114.421)				
Leibig *et al.* 2022 (27)	DM	Sistema IA propio (*in-house*) comparado con lectura independiente por 2 especialistas en radiología, para NT@97 %+SN@98 %	**AUC:** IA AUC: 0,982 (0,978-0,985)	*Sistema IA:* 89,8 % (88,5-91,1 %) *Dos especialistas en radiología lectura independiente:* 87,2 % (85,6-88,7 %)	2,6 % P < 0,0001	*Sistema IA:* 94,3 % (94,2-94,5 %) *2 especialistas en radiología lectura independiente:* 93,4 % (93,2-93,6 %)	1,0 % P < 0,0001

Tabla 10. **Tabla de contingencia 2×2 de los estudios individuales incluidos en el informe**

Estudio	Prueba índice (proveedor) / comparador	Casos	Cáncer	IA (*stand-alone*) (MD)					
				Sensibilidad	Especificidad	TP	FP	FN	TN
IA autónoma									
Hsu *et al.* (26)	IA *in-house*.	121.753	723	0,547	0,93	395	8.472	328	112.558
	IA *in-house*.			0,826	0,697	597	36.672	126	84.358
	Comparador. Lector único			0,826	0,93	597	8.472	126	112.558
Leibig *et al.* (27)	IA propia (*in-house*).	82.851	2.793	0,846	0,913	2.363	6.965	430	73.093
	Comparador. Doble lectura			0,872	0,934	2.435	5.284	358	74.774
Romero-Martin *et al.* (28)	IA (Transpara v 1.7.0). Punto de corte 80	15.999	113	0,628	Ne	71	—	42	—
	IA (Transpara v 1.7.0). Punto de corte 74			0,708	Ne	80	—	33	—
	Comparador. Lector único			0,584	Ne	66	—	47	—
	Comparador. Doble lectura			0,673	Ne	76	—	37	—
Sharma *et al.* (29)	IA (Mia v 2.0.1) (Muestra 10 años)	275.900	2.792	0,781	0,912	2.181	24.034	611	249.074
	IA (Mia v 2.0.1) (Muestra un año)			0,752	0,914	2.100	23.487	692	249.621
	Comparador. Lector único (Muestra 10 años)			0,764	0,96	2.133	10.924	659	262.184
	Comparador. Lector único (Muestra 1 año)			0,701	0,966	1.957	9.286	835	263.822

…/…

.../...

Estudio	Prueba índice (proveedor) / comparador	Casos	Cáncer	IA (*stand-alone*) (MD)					
				Sensibilidad	Especificidad	TP	FP	FN	TN
IA como ayuda al lector									
Bao *et al.* (21)	IA (Yizhun v 3.2.3)	560	345	0,9507	0,5488	331	97	17	117
	Comparador. Lector único (media)			0,8477	0,6355	292	78	53	137
Dang *et al.* (22)	IA (Mammoscreen v 1.2)	628	128	0,7	0,81	90	95	38	405
	Comparador. Lector único (media)			0,66	0,79	84	105	44	395
Lee *et al.* (23)	IA (Lunit v 1.1.1.0)	200	140	0,83	0,672	116	20	24	40
	Comparador. Lector único (media)			0,629	0,687	88	19	52	41
Sun *et al.* (24)	IA *in-house*	200	70	0,6878	0,8834	48	15	22	115
	Comparador. Lector único (media)			0,687	0,8205	48	23	22	107
vanWinkel *et al.* (DBT) (25)	IA (Transpara 1.6.0)	480	146	0,792	Ne	116	—	30	—
	Comparador. Lector único (media)			0,746	Ne	109	—	37	—
IA para clasificación									
Larsen *et al.* (30)	IA (Transpara 1,7,0)								
	Puntuación de la IA = 10	122.969	957	0,778	0,905	745	11.638	212	110.374
	Puntuación de la IA > 9,13	122.969	957	0,758	0,918	725	10.064	232	111.948
	Puntuación de la IA > 9,43	122.969	957	0,695	0,947	665	6.471	292	115.541
Lauritzen *et al.* (31)	IA (Transpara 1,7,0)								
	Puntuación de la IA ≤ 5 y UR = 9,989	114.421	1.118	0,697	0,986	779	1.586	339	111.717
Leibig *et al.* (27)	IA *in-house*								
	NT@97 %+SN@98 %	82.851	2.793	0,898	0,943	2.508	4.563	285	75.495

II.3.2. Resultados pregunta de investigación 2

¿Es coste-efectivo el uso de sistemas de IA en el cribado mamográfico de cáncer de mama en mujeres participantes en los PDPCM en comparación con la estrategia de cribado habitual realizada en los PDPCM?

II.3.2.1. Resultados de la búsqueda bibliográfica

La búsqueda bibliográfica realizada en las bases de datos electrónicas identificó 898 estudios como potencialmente relevantes. Una vez eliminadas las referencias duplicadas, se identificaron 646 para su lectura por título y resumen. Excluidos aquellos que no cumplieron con los criterios de inclusión, se seleccionaron dos referencias para su lectura a texto completo. De estas se seleccionó una para el análisis de su calidad y síntesis de la evidencia.

Los estudios excluidos tras la lectura a texto completo y las razones de su exclusión se recogen en el Anexo VI.3.3.

En la figura 5 se muestra el diagrama de flujo que resume el proceso de selección de estudios para responder a la pregunta de investigación.

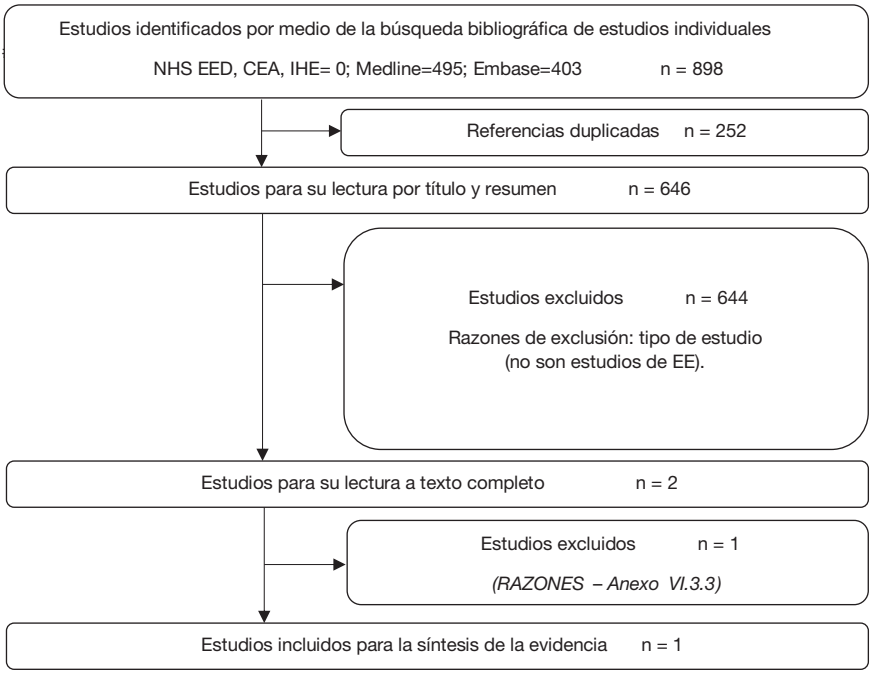


Figura 5. **Diagrama de flujo del proceso de selección de los estudios**

II.3.2.2. Descripción del estudio incluido

La búsqueda bibliográfica proporcionó un estudio (44) de EE para el análisis y síntesis de la evidencia. Este estudio se realizó en EE.UU. con el objetivo de examinar la relación coste-efectividad del uso de la IA o de la puntuación de riesgo poligénico (PRS) para guiar el cribado mamográfico del cáncer de mama en comparación con el cribado basado exclusivamente en la historia familiar (similar a las directrices del U.S. Preventive Services Task Force (USPSTF)), el cribado anual para todas las mujeres (similar a las directrices del American College of obstetricians and Gynecologists (ACOG) y la American College of Radiology (ACR)) y ningún cribado. El estudio simuló una cohorte de 100.000 mujeres blancas de 40 años sin historia previa de cáncer de mama. Cada mujer fue clasificada en tres categorías de riesgo subyacente de desarrollar cáncer de mama basado en una distribución de riesgo estimada para mujeres blancas de EE.UU. (45): (i) «verdadero» bajo riesgo definido como aquellas mujeres con un riesgo subyacente de cáncer de mama < 1,1 veces el riesgo medio en la población de mujeres de 40 años (riesgo relativo (RR) < 1,1), (ii) «verdadero» alto riesgo definido como aquellas con un RR > 1,1 y < 4, y (iii) «verdadero» muy alto riesgo definido como aquellas con un RR ≥ 4.

En el estudio se compararon ocho estrategias de cribado: (1) «no cribado» a ninguna edad independientemente del nivel de riesgo; (2) «cribado anual para todas las mujeres» a partir de los 40 años (similar al recomendado por la ACOG y la ACR) independientemente del nivel de riesgo; (3) «IA + no cribado para bajo riesgo»; (4) «IA + cribado bienal para bajo riesgo»; (5) «PRS + no cribado para bajo riesgo»; (6) «PRS + cribado bienal para bajo riesgo»; (7) «historia familiar + no cribado para bajo riesgo»; y, (8) «historia familiar + cribado bienal para bajo riesgo». Las estrategias 3 a 8 se diferenciaron en los enfoques adoptados para la predicción del riesgo y en las distintas frecuencias de cribado entre mujeres de bajo riesgo de 40 a 49 años. Así, para predecir el riesgo de cáncer de mama en las estrategias de cribado 3 y 4 se realizó una mamografía índice a los 40 años interpretada con IA para predecir el riesgo de cáncer de mama; en las estrategias 5 y 6 las mujeres se sometieron a pruebas genéticas en la que se genotipificaron 76 polimorfismos de un solo nucleótido que se saben asociados con el cáncer de mama; y en las estrategias 7 y 8 el cribado se guió de acuerdo con la historia familiar de cáncer de mama a los 40 años (similar a las recomendaciones realizadas por la USPSTF). En las estrategias 3 a 6, las mujeres con riesgo elevado (RR ≥ 1,1) se sometieron a una mamografía digital anual desde los 40 años mientras que a las mujeres con riesgo bajo no se les realizó cribado o se les determinó cribado bienal. En la estrategia 7 se consideró que las mujeres menores de 50 años sin historia familiar no fueron

cribadas mientras que en la estrategia 8 fueron cribadas bienalmente. A partir de los 50 años las estrategias de cribado 3 a 8 siguieron las directrices del USPSTF, es decir, las mujeres sin historia familiar de cáncer de mama (riesgo bajo) fueron cribadas bienalmente mientras las mujeres con historia familiar (riesgo alto) fueron cribadas anualmente. En todas las estrategias el cribado finalizó a los 74 años.

Para estimar los costes y la efectividad de las ocho estrategias de cribado se llevó a cabo un modelo híbrido de árbol de decisión/microsimulación. El análisis se condujo desde la perspectiva del sistema sanitario, siendo la duración de los ciclos del modelo de un año y el horizonte temporal para toda la vida de las mujeres.

El componente de árbol de decisión del modelo captó la predicción y estratificación del riesgo a los 40 años en función de la IA, la PRS o la historia familiar. Al entrar en el modelo las mujeres tenían un riesgo subyacente «verdadero» bajo, alto o muy alto riesgo de cáncer de mama. Las mujeres con riesgo muy alto (RR ≥ 4) se sometieron a cribado anual independientemente de la estrategia de cribado (excepto en la estrategia «Sin cribado»). Para las restantes, se determinó la medida en que la categoría de riesgo estimada coincidía con la categoría de riesgo subyacente en función de la precisión de cada estrategia de estratificación del riesgo (IA, ERP o historia familia).

El componente de microsimulación simuló el cribado, diagnóstico, progresión de la enfermedad y mortalidad para cáncer de mama. Todas las mujeres que entraron en el modelo no tenían cáncer, pero podrían desarrollar un cáncer *in situ* o invasivo a lo largo del tiempo en función de las tasas de incidencia por edad observadas, pudiendo el cáncer *in situ* posteriormente progresar a cáncer invasivo. Los cánceres invasivos se clasificaron en estadios locales, regionales y distantes. Las mujeres sometidas al cribado mamográfico tenían una mayor probabilidad de ser diagnosticadas con cáncer *in situ.* Un cribado mamográfico más agresivo también se tradujo en un mayor número de cánceres diagnosticados en estadios más tempranos (en lugar de más avanzados). Las mujeres que desarrollaron cáncer de mama invasivo se enfrentaron al riesgo de muerte por cáncer o por otras causas.

En el modelo, la precisión de la predicción del riesgo de cáncer de mama utilizando IA y PRS se midió utilizando el parámetro AUC-ROC obtenido de estudios publicados (46, 47). El valor del AUC-ROC utilizado fue de 0,71 para la IA (47) y de 0,69 para la PRS (46). Para simular la distribución del RR estimado utilizando estos valores se siguió un método previamente publicado en dos estudios (48, 49). Las mujeres con un RR es-

timado ≥ 1,1 fueron clasificadas de alto riesgo y con un RR estimado < 1,1 de bajo riesgo. Dado que el valor AUC-ROC para IA y PRC fue < 1, no todas las mujeres con 'verdadero' alto riesgo fueron correctamente clasificadas como tales.

En las estrategias en las que la predicción del riesgo se basó en la historia familiar, dado que las mujeres con bajo riesgo subyacente no tenían una historia familiar de cáncer de mama, todas las mujeres con bajo riesgo fueron correctamente clasificadas como tales. Entre las mujeres con alto riesgo, se asumió que el 37 % serían correctamente clasificadas, proporción calculada como la parte de mujeres estadounidenses con historia familiar de primer grado de cáncer de mama (16 %) entre las mujeres de alto riesgo (43 % de la cohorte a estudio).

Para estimar la probabilidad de desarrollar cáncer de mama *in situ* o invasivo, se multiplicó la tasa de incidencia de cáncer de mama específica por edad por 100.000 mujeres blancas estadounidenses (ajustada por el aumento en la tasa de incidencia debido al cribado mamográfico) por el RR «verdadero» de las mujeres. El estadio en el momento de la detección del cáncer dependía de la frecuencia del cribado y de la sensibilidad de la mamografía; esta última dependía de la edad de la paciente y se obtuvo de la literatura publicada (50). La distribución del estadio en el diagnóstico en ausencia de cribado se calculó con base en la proporción de cánceres locales, regionales y distantes observados entre las mujeres blancas ≤ 50 años durante 1975-1979. La distribución de estadios con cribado anual o bienal se obtuvo de estimaciones basadas en datos entre 1996-2012 del *Breast Cancer Surveillance Consortium*. Las pacientes diagnosticadas de cáncer de mama invasivo se enfrentaban al riesgo de mortalidad por cáncer de mama hasta 20 años después del diagnóstico. Este riesgo era específico a la edad, estadio en el diagnóstico, así como al estado del receptor de estrógeno y del factor de crecimiento epidérmico humano. Todas las mujeres se enfrentaron a un riesgo de mortalidad por causas no relacionadas con el cáncer de mama específico por edad, que se obtuvo restando la mortalidad específica por edad por cáncer de mama de las tablas de vida estadounidenses de 2017.

Para cada estrategia se incluyó el coste de la predicción del riesgo, el coste del cribado con DM y el coste del tratamiento del cáncer de mama que depende del estado en el que el cáncer se diagnostique (menor coste para los detectados en estado más temprano). El coste de las pruebas genéticas para determinar el PRS fue el coste de la prueba OncoArray en laboratorios estadounidenses. Además, se asumió que las pacientes recibieron asesoramiento genético antes y después de las pruebas genéticas. El coste de la predicción basada en IA se obtuvo de una publicación de la So-

ciedad Europea de Radiología (51). El coste de la mamografía se obtuvo de la Lista de Honorarios Médicos 2020 del Centro de Medicare y Medicaid. El coste de las pruebas diagnósticas tras un diagnóstico positivo y el coste del tratamiento de cáncer de mama se obtuvo de la literatura publicada (52, 53). Todos los costes se estimaron en dólares estadounidenses ($) de 2020 y se descontaron al 3 % por año.

La efectividad se midió en años de vida ajustados por calidad (AVAC). El cribado supuso una desutilidad de 0,006 AVAC durante una semana, y el estudio diagnóstico, tras un resultado de cribado positivo, supuso una desutilidad de 0,105 AVAC durante cinco semanas (52). Las utilidades fueron específicas para la edad y estadio de cáncer. Para todos los estadios del cáncer, la utilidad en el primer año después del diagnóstico de cáncer fue menor que en años posteriores. Todos los valores de utilidad se descontaron al 3 % por año.

Se estimó el coste y la efectividad para las ocho estrategias. Una estrategia fue considerada coste-efectiva frente a otra si el RCEI fue menor que el umbral de disponibilidad a pagar de 100.000 $ por AVAC ganado. Una estrategia fue dominada si fue más costosa y menos efectiva que otra estrategia y dominada por extensión si conseguía menos AVAC totales que una estrategia más costosa a un coste incremental por AVAC más elevado, es decir, su RCEI relativa a la siguiente estrategia menos costosa era superior a la RCEI de una estrategia más eficaz.

Se realizó un análisis que incluía cuatro estrategias adicionales similares a las estrategias 3 a 6, excepto que la predicción del riesgo se realizó exclusivamente utilizando IA o PRS (no se consideraron factores de riesgo personales y demográficos). El valor AUC-ROC para estas cuatro estrategias fue de 0,69 para la IA y de 0,63 para PRS.

Se llevó a cabo análisis de sensibilidad univariante variando los valores de los costes y utilidades clave y se abordó la incertidumbre de los parámetros mediante un análisis de sensibilidad probabilístico. A continuación, se varió el AUC-ROC original de la IA y PRS en un ± 20 %.

También se examinó la solidez de los resultados en función de la elección del umbral de RR utilizado para definir el alto riesgo estimado. Siguiendo estudios anteriores, se utilizaron umbrales alternativos de 1,3 y 2 (en lugar del 1,1 utilizado en el análisis del caso base).

Por último, se validó el modelo mediante un análisis en el que se comparó la incidencia acumulada de cáncer de mama a lo largo de la vida modelizada y la mortalidad con cribado con las proporciones observadas recientemente. A continuación, se validaron de forma cruzada las estima-

ciones incrementales de los costes, los AVAC y las tasas de FP (en comparación con la ausencia de cribado) frente a estudios previos para la estrategia en la que la predicción del riesgo se basa en los antecedentes familiares y las personas sin antecedentes familiares se someten a un cribado bienal a partir de los 50 años.

II.3.2.3. Calidad de la evidencia del estudio incluido

La calidad de la evidencia del estudio de Mital *el al.* (44), único estudio de EE incluido queda reflejada en la tabla 11. De acuerdo con la herramienta FLC 3.0, la calidad de la evidencia aportada por dicho estudio se evaluó como alta.

De las seis áreas de las que se compone la FLC, cinco fueron calificadas «sí» y una «parcialmente». El dominio validez externa se evaluó «parcialmente» porque se consideró que los resultados alcanzados en el estudio podrían no ser del todo generalizables al contexto del SNS analizado en esta revisión. Esto se debió a que las fuentes de datos utilizadas para la obtención de los costes, la efectividad y el resto de las variables necesarias para completar el modelo, se obtuvieron principalmente de la literatura y de bases de datos estadounidenses.

Tabla 11. **Valoración de la calidad**

Ref	Pregunta	Métodos	Resultados	Conclusiones	Conflicto de intereses	Validez externa	Calidad del estudio
Mital S *et al.* 2022 (44)	Sí	Sí	Sí	Sí	Sí	Parcialmente	Alta

II.3.2.4. Descripción y análisis de los resultados

El coste unitario de las pruebas genéticas, el de la predicción basada en IA, el de la mamografía, el de las pruebas diagnósticas tras un diagnóstico positivo y del tratamiento de cáncer de mama queda reflejado en la tabla 12.

Tabla 12. **Coste de la predicción del riesgo, del cribado con mamografía digital y del diagnóstico y tratamiento**

Coste	Valor en $ de 2020 (DS)	Fuente
IA	112 (28)	European Society of Radiology/cálculos propios
Prueba genética OncoArray	115 (29)	Iowa Institute of Human Genetics
Asesoramiento genético/sesión	44 (11)	Sun *et al.* (24)
Mamografía	152 (38)	Centers for Medicare and Medicaid Services
Diagnóstico (verdaderos positivos)		Stout *et al.* (52)
Edad 40-49	2.491 (623)	
Edad 50-64	2.337 (584)	
Edad 65-74	2.350 (588)	
Diagnóstico (falsos positivos)		Stout *et al.* (52)
Edad 40-49	261 (65)	
Edad 50-64	309 (77)	
Edad 65-74	310 (77)	
Tratamiento		Schousboe *et al.* (54) / Shih *et al.* (53)
In situ, coste inicial	11.543 (2.886); 10.329 (2.582)	
In situ, coste continuado	0	
Localizado, coste inicial	29.374 (7.343); 18.995 (4.749)	
Localizado, coste continuado	1.986 (497); 1.267 (317); 1.210 (303); 1.446 (362); 1.044 (261); 817 (204)	
Localizado, coste termina	51.800 (12.950)	
Regional, coste inicial	51.859 (12.965); 35.365 (8.841)	
Regional, coste continuado	6.747 (1.687); 4.572 (1.143); 4.315 (1.079); 3.744 (936); 2.662 (666); 2.353 (588)	
Localizado, coste terminal	58.172 (14.543)	
Distante, coste inicial	56.702 (14.176); 43.543 (10.886)	
Distante, coste continuado	23.581 (5.895); 20.945 (5.236); 20.162 (5.040); 17.744 (4.436); 13.094 (3.274); 13.478 (3.370)	
Distante, coste terminal	73.970 (18.493)	
Tamoxifen (5 años)	1.519 (76)	
Trastuzumab	81.717 (20.429)	

El cálculo del coste unitario de la lectura mamográfica mediante IA (112 $) se realizó con base al coste fijo (60.000 € (65.300 $)) y coste anual por la licencia del software (20.000 € (21.770 $)) propuesto por la Sociedad Europea de Radiología (51), a la amortización del equipo (10 años) y al número de mujeres de 40 años (2 millones) atendidas en los 8.695 centros mamográficos en EE.UU.

El coste inicial de tratamiento para cada estadio se calculó para mujeres < 70 años y para ≥ 70 años, respectivamente, como la media ponderada de los costes de los distintos tratamientos del cáncer de mama, con la proporción ponderada de pacientes que reciben cada tipo de tratamiento. Los costes del tratamiento continuado para cada estadio son de 1 a 5 y ≥ 6 años después del año del diagnóstico, respectivamente.

Los valores de desutilidad del cribado y del estudio diagnóstico tras un resultado de cribado positivo y de utilidad para los distintos estadios de salud utilizadas para el cálculo de los AVAC, quedan reflejadas en la tabla 13:

Tabla 13. **Valores de utilidad/desutilidad**

Utilidades	Valor (DS)	Fuente
Desutilidad del cribado	0,006 (0,00003) para una semana	Stout *et al.* (52)
Desutilidad del diagnóstico adicional	0,105 (0.00001) para cinco semanas	Stout *et al.* (52)
Estadios de salud*		Schousboe *et al.* (54)
Sano	0,762-0,859	
In situ	0,689-0,777	
Local	0,645-0,842	
Regional	0,574-0,777	
Distante	0,574-0,715	

* Los valores de utilidad para los estadios de salud dependen de la edad y del momento en el que comienza el diagnóstico.

En la tabla 14 se resumen los costes a lo largo de la vida, los AVACs ganados y los resultados de cáncer de mama con cada estrategia de cribado.

Tabla 14. **Costes a lo largo de la vida, AVACs ganados, resultados de cáncer de mama con cada estrategia de cribado**

Estrategia	Coste (en 1.000 $)	Efectividad (en AVACs)	N.º (%) de mujeres de alto riesgo real clasificadas como de alto riesgo	N.º (%) de mujeres de bajo riesgo real clasificadas como de bajo riesgo	N.º de muertes por cáncer de mama (por 100.000 mujeres)	N.º de diagnósticos FP
No cribado	1.745.808	1.976.720	—	—	3.367	0
Historia familiar + no criba para bajo riesgo	1.823.664	1.978.241	15.461 (36,0)	56.444 (100)	2.988	121.737
IA + no cribado para bajo riesgo	1.843.441	1.980.830	24.525 (57,0)	49.122 (87,0)	2.956	141.339
PRS + no cribado para bajo riesgo	1.852.227	1.980.713	24.381 (56,7)	48.805 (86,4)	2.936	141.443
Historia familiar + cribado bienal para bajo riesgo	1.879.254	1.980.731	15.461 (36,0)	56.444 (100)	2.916	170.917
PRS + cribado bienal para bajo riesgo	1.909.968	1.978.418	24.381 (56,7)	48.805 (86,4)	2.885	180.219
IA + cribado bienal para bajo riesgo	1.910.153	1.978.604	24.525 (57,0)	49.122 (87,0)	2.903	180.163
Cribado anual para todos	2.022.120	1.978.717	—	—	2778	290.325

La estrategia de «no cribado» presenta el menor coste (1.745.808 $ por 100.000 mujeres) y la estrategia «IA + no cribado para las mujeres de bajo riesgo» el mayor número de AVACs ganados (1.980.830 por 100.000 mujeres), situándose ambas en la frontera coste-efectividad. Observando los resultados obtenidos, la estrategia «IA + no cribado para las mujeres de bajo riesgo» domina por extensión a la estrategia «historia familiar + no cribado para las mujeres de bajo riesgo», es decir, es más efectiva y tiene un menor RCEI cuando se compara con la siguiente más barata («no cribado»), y domina a las otras 5, es decir, es más barata y efectiva.

«IA + no cribado para las mujeres de bajo riesgo» fue la estrategia más coste-efectiva. En comparación con la estrategia de «no cribado» costó 97.633 $ más por 100.000 mujeres, pero ocasionó 4.110 AVAC adicionales por 100.000 mujeres. El RCEI de la estrategia «IA + no cribado para las

mujeres de bajo riesgo» en comparación con «no cribado» fue de 23.755 $ por AVAC ganado, menor que el umbral de disponibilidad a pagar de 100.000 $ por AVAC ganado.

El análisis de sensibilidad univariante realizado indicó que el RCEI fue más sensible a los costes de la mamografía y a los costes y utilidades específicos del estadio de salud. Para todos los valores de esos costes y utilidades en el rango de ±25 %, «IA + no cribado para las mujeres de bajo riesgo» continuó siendo más coste-efectivo que «no cribado». Se observó que la estrategia «IA + no cribado para las mujeres de bajo riesgo» continuó siendo la más coste-efectiva siempre que el coste de la lectura con IA permaneciese por debajo de 318 $ por mamografía.

El análisis de sensibilidad probabilístico señaló que para un umbral de disponibilidad a pagar de 100.000 $ / AVAC ganado, la probabilidad de que la estrategia «IA + no cribado para las mujeres de bajo riesgo» fuese coste-efectiva fue del 96 %.

Los análisis de sensibilidad adicionales realizados, indicaron que la estrategia «IA + no cribado para las mujeres de bajo riesgo» fue la estrategia más coste-efectiva incluso cuando los valores de IA y PRS utilizados variaron un ±20 % del valor base.

El análisis realizado para umbrales de RR igual a 1,3 y 2 mostró que la estrategia «PRS + no cribado para las mujeres de bajo riesgo» generaba mayores AVACs pero también mayores costes que «IA + no cribado para las mujeres de bajo riesgo», dando lugar a un RCEI superior al umbral de disponibilidad a pagar de 100.000 $ / AVAC ganado, por lo que «AI + no cribado para las mujeres de bajo riesgo» continuaba siendo la estrategia más óptima.

La evaluación de la validez del modelo indicó que los datos modelados sobre la incidencia y la mortalidad por cáncer de mama a lo largo de toda la vida de las mujeres fueron 16 % y 2,9 % con cribado, respectivamente, similares a las proporciones observadas para mujeres blancas entre 2016-18 según el Programa de Investigación sobre Vigilancia del Instituto Nacional de Cáncer (SEER), 13 % y 2,5 %, respectivamente. Además, el coste incremental y AVAC incremental estimados en el estudio para la estrategia «historia familiar + no cribado para las mujeres de bajo riesgo» fue similar al encontrado en un estudio reciente coste-efectividad de alta calidad (53), 778 $ vs 682 $ ($ del 2020) y 0,015 vs 0,017, respectivamente. Por último, el número de FP diagnosticados para la estrategia anterior, 121.737 por 100.000 mujeres entraba dentro del rango indicado por la USPSTF, 830-1.325 por 1.000 mujeres.

II.4. Discusión

II.4.1. Hallazgos principales

Para la pregunta de investigación *¿El uso de sistemas de IA integrados en los PDPCM para la detección de cáncer de mama, es más, igual o menos preciso en comparación con la estrategia de cribado habitual realizada en los PDPCM?* se han identificado y resumido los resultados de una RS, en la que se analizaron 12 artículos, y de 11 estudios individuales. Todos los estudios evaluaron la precisión de sistemas de IA de redes neuronales convolucionales, comercialmente disponibles o propios, para informar sobre la capacidad de estos sistemas para detectar correctamente a las mujeres con cáncer (sensibilidad) y sin cáncer (especificidad). En ellos, la IA jugó un papel en el proceso de cribado de mama como sistema autónomo, como ayuda al especialista en radiología y como herramienta de clasificación previa al cribado.

Como sistema de lectura autónomo, nueve estudios (26-29, 32-36) (681.753 casos, 8.836 cánceres) analizaron la sensibilidad y especificidad de los sistemas de IA. En cinco estudios (280.806 casos, 4.556 cánceres) la capacidad discriminativa de los sistemas de IA se consideró buena (AUC-ROC ≥ 0,9) y en tres (125.049 casos, 1.488 cánceres) aceptable (AUC-ROC ≥ 0,8 y < 0,9). En los tres estudios de menor tamaño (3.581 casos, 896 cánceres) y en el de mayor tamaño (275.900 casos, 2.792 cánceres) la sensibilidad de los sistemas de IA para la detección de cánceres de mama fue superior a la de la lectura original única del especialista en radiología para un umbral de predicción de la IA igual a la especificidad media del primer lector (del 67 % y del 79 %) o para un umbral preespecificado de revaloración o no. En un estudio (113.663 casos, 793 cánceres), solo para uno de los tres sistemas de IA analizados se señaló una sensibilidad superior a la de la lectura del primer y segundo lector, aunque inferior a la lectura de consenso, para una especificidad igual a la del primer lector del 96,6 %. En otro (15.999 casos, 113 cánceres), la sensibilidad fue no inferior a de la lectura radiológica única y doble para unos puntos de corte del sistema de IA igual a 80 y 74. Por el contrario, en dos estudios (204.604 casos, 3.516 cánceres), la sensibilidad y la especificidad de la IA fue inferior a la de la lectura original del especialista en radiología, para un umbral, definido en uno de ellos, igual a una sensibilidad y especificidad del especialista en radiología del 83 % y 93 %, respectivamente. Por último, en otro estudio (68.008 casos, 780 cánceres) la especificidad de la IA fue inferior a la del primer lector original y a la de la lectura original de consenso para una sensibilidad igual a la del primer lector original del 77 %.

Como ayuda a la lectura radiológica, ocho estudios (21-25, 37-39) (2.670 casos, 1.139 cánceres) analizaron la sensibilidad y especificidad de los sistemas de IA; siete examinaron DM y uno DBT. En los siete estudios (2.190 casos, 993 cánceres) en donde se analizaron DM, los sistemas de IA mostraron una capacidad discriminativa buena en uno (AUC-ROC ≥ 0,9), aceptable en cuatro (AUC-ROC ≥ 0,8 y < 0,9) y no aceptable en dos (AUC-ROC < 0,8). Cuando se comparó la capacidad discriminativa de la lectura radiológica con ayuda de la IA frente a sin ayuda de la IA, se observó que fue superior para la primera, siendo la diferencia estadísticamente significativa en seis de ellos. Además, un estudio (480 casos, 146 cánceres) que analizó mamografías DBT alcanzó un valor AUC-ROC aceptable, siendo la diferencia cuando se comparó con IA sin ayuda estadísticamente significativa. En los siete estudios en donde se analizaron DM, la sensibilidad evaluada como la media de la lectura realizada por especialistas en radiología con la ayuda de los sistemas de IA, fue superior a la realizada sin ayuda de sistemas de IA. En cinco (1.362 casos, 795 cánceres) la diferencia fue estadísticamente significativa y en dos (828 casos, 198 cánceres) no. Con lo que respecta a la especificidad, en cuatro estudios (1.308 casos, 438 cánceres) fue superior para la lectura radiológica con apoyo de la IA en comparación a la realizada sin ayuda, y en tres (882 casos, 575 cánceres) fue inferior. Las diferencias encontradas solo fueron estadísticamente significativas para un estudio en el que la especificidad fue menor con ayuda que sin ayuda. En el estudio en el que se evaluaron DBT, la sensibilidad de la lectura radiológica con ayuda de la IA fue mayor que sin ayuda, diferencia estadísticamente significativa, y la especificidad menor.

Como herramienta de clasificación previa al cribado, siete estudios (27, 30, 31, 40-43) (439.033 casos, 5.510 cánceres) evaluaron la sensibilidad y especificidad de los sistemas de IA utilizando distintos puntos de corte para señalar la probabilidad de cáncer visible en la mamografía. En cinco estudios (201.643 casos, 3.435 cánceres), para umbrales bajos, la sensibilidad fue alta, variando entre el 88,5 % y el 97,4 %. En los otros 2 (237.390 casos, 2.075 cánceres), la sensibilidad varió entre el 70 % y el 78 %, para umbrales de puntuación altos.

En siete estudios (21-24, 37, 38, 43) en los que el sistema de IA se utilizó como ayuda al especialista en radiología y en dos estudios (25, 39) en los que se utilizó como herramienta de clasificación, se analizó el efecto de los sistemas de IA en **la carga de trabajo de los especialistas en radiología**. En los siete estudios en los que se utilizó como apoyo a la lectura, la carga de trabajo se midió como el tiempo de lectura que emplearon los especialistas en radiología con y sin ayuda de la IA, el cual fue recogido automáticamente por el software empleado. Los resultados obtenidos variaron de

unos estudios a otros. En cuatro se observó una disminución del tiempo de lectura con ayuda de la IA, y en tres un aumento. Además, en tres estudios se analizó la disminución del tiempo de lectura para las puntuaciones de riesgo de sospecha de cáncer de mama, señalándose que fue menor para exámenes de baja sospecha, puntuación de 1 a 5, y mayor en los de sospecha alta, puntuación de 6 a 10. En los dos estudios en los que la IA se empleó para la clasificación la carga de trabajo se evaluó en función del número de lecturas realizadas por los especialistas en radiología, siendo menor la carga de trabajo cuando se utilizaron los sistemas de IA.

La evidencia analizada indica que la capacidad discriminativa para detectar pacientes con cáncer frente a pacientes sin cáncer de los sistemas de IA utilizados como sistema de lectura autónomo o como ayuda a la lectura radiológica es buena o aceptable en la mayoría de los algoritmos analizados. Además, señala que los sistemas de IA utilizados como ayuda a la lectura o como herramienta de clasificación previa al cribado identifican con una precisión adecuada a las mujeres con cáncer de mama, no siendo así cuando se utilizan como sistema de lectura autónomo. Por último, se observa una posible disminución en la carga de trabajo de los especialistas en radiología cuando los sistemas de IA se emplean como ayuda a la lectura o como herramienta de clasificación previa al cribado.

Los estudios incluidos en esta RS para su análisis muestran heterogeneidad en la metodología, lo que puede suscitar cierta preocupación por el riesgo de sesgo y aplicabilidad que pueden presentar. Los estudios con diseño retrospectivo son propensos al sesgo de verificación parcial. Este surge porque se utilizan dos estándares de referencia diferentes para el resultado con el fin de verificar la enfermedad de interés en diferentes grupos de pacientes: *1er estándar de referencia)* diagnóstico de cáncer en la evaluación (mediante prueba de biopsia); *2º estándar de referencia)* seguimiento y diagnóstico de cáncer sintomático (cáncer de intervalo) o en el siguiente cribado. Este sesgo no permite cuantificar el efecto global sobre la precisión o los cánceres de intervalo cuando la IA se integra en la vía de cribado, ya que no pueden medir el efecto de la IA cuando se incorpora al flujo de trabajo con los especialistas en radiología, ni establecer correctamente la precisión de la prueba o el espectro de enfermedades detectadas si la concordancia entre la lectura de la IA y la de los especialistas en radiología es baja o si la IA tiene una sensibilidad sustancialmente mayor (6). Por otro lado, una parte de los estudios analizados adolecen de un sesgo de selección de pacientes al utilizar cohortes de estudio de lectores enriquecidos o muestras de conveniencia en lugar de muestras consecutivas en entornos reales de exámenes de cribado. Los estudios de laboratorio enriquecidos, con múltiples lectores, múltiples casos y conjuntos de pruebas

ocasionan el llamado efecto laboratorio debido al cual la precisión de los especialistas en radiología independientes en estos estudios no puede generalizarse a la práctica clínica como consecuencia de las diferencias existentes en las condiciones de lectura y la prevalencia del cáncer entre el laboratorio y la práctica clínica. Esto produce una baja generalidad para la práctica clínica. No se puede evaluar ni comparar el efecto de las decisiones de la IA con las de los especialistas en radiología para resultados clínicamente significativos, debido a la dificultad de extrapolar el cambio completo de la vía (es decir, utilizando IA) en la práctica clínica, donde la prevalencia del cáncer es menor, a partir de la única decisión tomada en un entorno de laboratorio (6).

Hubo estudios en los que no se establecieron umbrales de positividad clínicamente relevantes o si se hizo variaron entre los estudios analizados. Esto puede implicar que una modificación del umbral de positividad aumente la sensibilidad disminuyendo la especificidad. La determinación del umbral más apropiado es un aspecto fundamental de la evaluación de la precisión de las pruebas ya que induce un equilibrio entre sensibilidad y especificidad. Como señalan Taylor-Philips *et al.* (6), análisis con umbrales clínicamente relevantes son más útiles que analizar el valor AUC-ROC, porque la forma de la curva ROC y su valor AUC no afectan a los resultados clínicos, sólo la precisión de la prueba en el umbral utilizado en la práctica es relevante para los resultados de las mujeres en el cribado.

Otro punto a tener en cuenta es que, tal y como lo señalaron Freeman *et al.* (5), no se pudo evaluar el espectro del efecto resultante de los estudios porque estos no informaron adecuadamente de la distribución de los hallazgos radiológicos originales, como la distribución de las puntuaciones BI-RADS originales. Sin embargo, señalaron que era probable que el efecto fuera mayor cuando la selección se basó en características de la imagen o del cáncer que si el enriquecimiento se conseguía incluyendo a todas las mujeres con cáncer disponible y una muestra aleatoria de las que eran negativas.

Por último, la definición del estándar de referencia para la definición de casos normales presenta ciertas variaciones entre los estudios incluidos. Así, los casos normales se definieron mediante una simple decisión consensuada de los especialistas en radiología o como las mujeres con resultado negativo para un periodo de seguimiento de uno, dos o tres años. Esta inconsistencia significó que las estimaciones de precisión fueron comparables en, pero no entre estudios (5). Además, que el periodo se seguimiento fuese inferior a dos años en algunos estudios pudo dar lugar a una subestimación del número de cánceres no detectados y a una sobreestimación de la precisión de la prueba.

Para responder a la pregunta *¿Es coste-efectivo el uso de sistemas de IA en el cribado mamográfico de cáncer de mama en mujeres participantes en los PDPCM en comparación con la estrategia de cribado habitual realizada en los PDPCM?* se ha identificado y resumido un solo estudio de EE (44) en el que se examinó la relación coste-efectividad del uso de la IA o de la puntuación PRS para guiar el cribado mamográfico del cáncer de mama en el grupo de mujeres de 40-49 años y sin historia previa de cáncer, en comparación con el cribado basado en la historia familiar, el cribado anual para todas las mujeres o el no cribado.

El estudio indicó que para el grupo de mujeres entre 40-49 años, la estrategia en la que se utilizó la IA para la predicción del riesgo de cáncer de mama seguida de no cribado para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF, fue la más coste efectiva en comparación con el resto de estrategias analizadas: no cribado, cribado anual para todas las mujeres, cribado mediante IA seguida de cribado bienal para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF, cribado mediante PRS seguido de no cribado para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF, cribado mediante PRS seguido de cribado bienal para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF, cribado guiado de acuerdo con la historia familiar de cáncer de mama seguido de no cribado para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF y cribado guiado de acuerdo con la historia familiar de cáncer de mama seguido de cribado bienal para las mujeres de bajo riesgo hasta los 74 de acuerdo con los criterios de la USPSTF.

Los análisis de sensibilidad determinístico y probabilístico realizados en el estudio, así como los llevados a cabo para examinar la validez de los resultados y del modelo, indicaron que para el grupo de mujeres entre 40-49 años, la estrategia en la que se utilizó la IA para la predicción del riesgo de cáncer de mama seguida de no cribado para las mujeres de bajo riesgo más cribado de mama a partir de los 50 años hasta los 74 de acuerdo con los criterios de la USPSTF siguió siendo la más coste efectiva.

El resultado obtenido en el estudio analizado puede presentar problemas de aplicabilidad en un contexto diferente al evaluado debido a las fuentes de donde se extrajeron los datos utilizados para el cálculo de los costes, utilidades y demás parámetros necesarios para la cumplimentación del modelo. No se encontraron ensayos controlados aleatorizados en donde se comparó la IA con PRS o con otros criterios de cribado existen-

tes, por lo que la eficacia de la IA y PRS se tuvo que obtener de diferentes estudios. Los costes de las pruebas y de los diferentes estadios de salud se obtuvieron de bases de datos de precios y tarifas estadounidenses y de literatura. Además, al no conocerse el coste de la utilización de la IA para la predicción del riesgo de cáncer de mama en la práctica clínica se tuvo que estimar con base en datos publicados por la Sociedad Europea de Radiología.

II.4.2. Fortalezas y limitaciones

Como puntos fuertes de esta RS cabe señalar que se llevó a cabo una búsqueda bibliográfica exhaustiva realizada para identificar todos los estudios pertinentes, que se realizó una evaluación rigurosa del riesgo de sesgo de los estudios incluidos mediante la herramienta QUADAS-2 adaptada al objeto de la revisión, que la selección de estudios y la extracción de datos se realizó por duplicado, y que se emplearon unos criterios de inclusión suficientemente restrictivos para que solo se incluyeran estudios en los que los sistemas de IA analizados hubieran sido validados externamente en conjuntos de datos de diferentes centros del mismo país o de países diferentes de aquellos con los que se realizaron la evaluación interna. La validación interna sobreestima la precisión y puede dar lugar a una generalidad limitada de los resultados (55). Además, puede ocasionar sobreajuste y a una pérdida de generalidad debido a que el modelo se ajusta adecuadamente a los datos entrenados, pero en detrimento de su capacidad para funcionar con nuevos datos (56).

Como limitaciones cabe señalar que en la RS solo se incluyeron estudios publicados en inglés por lo que pudo existir evidencia relevante que no se haya incluido. Asimismo, en los estudios en los que se evaluaron múltiples algoritmos solo se analizó el que proporcionó el mejor rendimiento, por lo que se podría observar un pequeño sesgo hacia los sistemas de IA seleccionados. También se excluyeron aquellos estudios en los que los sistemas de detección asistida por ordenador para el cribado de mama fueron clasificados como tradicionales, lo que pudo haber excluido estudios relevantes que informaron mal de los sistemas de IA o que utilizaron una combinación de sistemas. Además, al extraerse de los sistemas de IA solo clasificaciones binarias, no se pudo conocer como afectaba a la realización de las pruebas de seguimiento otro tipo de información incluida en el formulario de evaluación de un especialista en radiología, como las características mamográficas o las puntuaciones BI-RADS de nivel de sospecha. Por último, señalar que los estudios no detallaron el impacto del formato de presentación de los resultados de la IA en la interpretación del especialista

en radiología y que en ningún estudio se analizó la aceptabilidad de la IA desde la perspectiva médico-legal o ética.

II.4.3. Acuerdos y desacuerdos con otras revisiones

En esta revisión se analizan más estudios y un mayor conjunto de datos (casos y cánceres) que los incluidos en RS anteriores similares (5, 8, 57) realizadas también con el objetivo de analizar la precisión de la IA en la detección de cáncer de mama.

Los resultados presentados en las revisiones sistemáticas realizadas por Freeman *et al.* (5), Hickman *et al.* (8) y Anderson *et al.* (57) son análogos a los obtenidos en la nuestra. Así, Freeman *et al.* (5) señalaron que cuando la IA se utilizó como sistema autónomo en sustitución de los especialistas en radiología, las estimaciones puntuales de la precisión de los sistemas de IA fueron inferiores a las obtenidas por consenso de dos especialistas en radiología en la práctica del cribado, con resultados dispares en comparación con un único especialista en radiología. Cuando se empleó como un sistema de clasificación indicaron que, a umbrales bajos, la IA podría alcanzar una sensibilidad alta por lo que podría ser adecuada como preselección para identificar a las mujeres de bajo riesgo cuyas mamografías requerían menos o ninguna revisión radiológica. Por último, cuando analizaron la IA como ayuda al lector observaron que las estimaciones puntuales de la sensibilidad media fueron mayores para los especialistas en radiología con apoyo de IA que para la lectura sin ayuda y que el efecto en la especificidad fue pequeño.

Al igual que en la revisión anterior, Anderson *et al.* (57) alcanzaron resultados similares sobre la predicción de los sistemas de IA en la detección de cáncer de mama cuando estos se utilizaron como sistema autónomo o como ayuda al lector. Los resultados indicaron que unos valores de AUC-ROC entre buenos y no aceptables para la IA como sistema autónomo y aceptables para la IA como ayuda. Cuando se compararon con el valor AUC-ROC de los especialistas en radiología, el AUC-ROC de lectura del especialista en radiología con ayuda de la IA fue mayor que sin ayuda, siendo las diferencias estadísticamente significativas. Sin embargo, cuando se comparó el valor AUC-ROC de la lectura radiológica frente a la IA como sistema autónomo los resultados fueron dispares, en algunos casos el rendimiento de la IA fue superior en otros significativamente peor. Además, la precisión de los sistemas de IA como sistema autónomo mostraron resultados inconsistentes cuando se comparó con la del especialista en radiología, mostrando una precisión mayor o menor o una especificidad menor con ganancias relativamente menores en la sensibilidad. Cuando se

analizó la lectura del especialista en radiología con ayuda de la IA frente a sin ayuda, sistemáticamente se observó un aumento de la precisión de los especialistas en radiología con ayuda, mayor sensibilidad y especificidad.

Por su lado, Hickman *et al.* (8) analizaron algoritmos de detección asistida por ordenador de aprendizaje automático como sistemas autónomos o para clasificación. En la RS con MA realizada obtuvieron como resultados que cuando la IA se utilizó para clasificación el número de mamografías examinadas por los especialistas en radiología disminuyó, sin que se produjesen cambios perjudiciales cuando el rendimiento de los especialistas en radiología fue extrapolado en un flujo de trabajo de cribado adaptado, es decir, cuando se utilizó la IA solo para la lectura de los casos considerados normales como alternativa a la lectura simple o doble. Cuando evaluaron la precisión de la IA como sistema autónomo señalaron que los sistemas de IA podrían ser equivalentes con la lectura radiológica. El MA realizado para este caso obtuvo un mayor valor AUC-ROC para la IA que para los especialistas en radiología.

II.5. Conclusiones

La evidencia revisada sugiere que los sistemas de IA son más precisos cuando se emplean en la fase del proceso de cribado como ayuda a la lectura única radiológica para asistir al especialista en radiología en la interpretación y proporcionar otra evaluación de la misma interpretación mamográfica, así como herramienta de clasificación previa al cribado aplicada para eliminar todos los casos normales evidentes, es decir, los de bajo riesgo, que no requieren más revisión, y un cribado posterior del resto de los casos.

El estudio de EE analizado señala que, entre todas las estrategias evaluadas en él, la estrategia más coste efectiva es la que realiza a todas las mujeres de 40 años una mamografía índice (que puede o no ser parte de los servicios de cribado de mama), la que se interpreta mediante IA para predecir el riesgo de cáncer de mama. A las mujeres en las que se prevé un riesgo elevado de cáncer de mama (RR ≥ 1,1) se le hace a una mamografía digital anual a partir de los 40 años, mientras que a las que se prevé un riesgo bajo no se las criba. Este patrón de cribado se mantiene hasta los 49 años. A partir de los 50 años y hasta los 74 años, el cribado sigue las directrices del USPSTF, es decir, se criba bienalmente a las mujeres sin historia familiar de cáncer de mama y anualmente a las que tienen historia familiar de riesgo de cáncer de mama.

Para medir el efecto de la IA en la práctica clínica se requiere la realización de estudios prospectivos al ser el método menos sesgado para medir la precisión. Los estudios prospectivos reducen o eliminan el sesgo de verificación parcial y diferencial al aplicar el mismo patrón de referencia a las mujeres clasificadas con resultados positivos tanto por la IA como por los especialistas en radiología, reducen o eliminan el sesgo de incorporación si ciegan las pruebas de seguimiento al tipo de pruebas índice, y pueden evaluar la interacción entre la IA y los especialistas en radiología.

Para conocer el efecto de la IA en los resultados de las mujeres, como parte del equilibrio entre los perjuicios y beneficios del cribado, se deben realizar ensayos controlados aleatorizados, prospectivos, de prueba-tratamiento, que asignen aleatoriamente a las mujeres a la vía original (lectura mamográfica por especialista en radiología única o lectura doble más consenso) o a la nueva vía propuesta (lectura mamográfica mediante sistemas de IA como ayuda al lector, sistema autónomo o para clasificación) y con un seguimiento suficiente que permita obtener resultados clínicamente significativos. Este tipo de estudios permiten aportar la evidencia necesaria para conocer el tipo, el estadio, las características de los cánceres detectados por la IA, lo que posibilita evaluar cambios potenciales en el balance beneficios daños. También proporciona evidencia para subgrupos de mujeres de acuerdo con la edad, la densidad mamaria, cáncer de mama previo e implantes mamarios.

La incorporación de la IA en los PDPCM precisa medir el conjunto de cambios que puede implicar en todo el proceso de cribado y no solo el efecto de la prueba por sí sola. Se necesita conocer el efecto de la utilización de la IA sobre el comportamiento y la aceptación de los especialistas en radiología, sobre su carga de trabajo y sobre las cuestiones éticas que puedan planteárseles. Además, es necesario comparar directamente diferentes sistemas de IA y evaluar la precisión de estos en un mismo mamógrafo y dentro de un programa, conocer el efecto de las diferencias existentes entre los PDPCM sobre la detección del cáncer con IA, la integración entre los sistemas de IA y los sistemas de información y gestión de los PDPCM y el efecto de poner a disposición de los sistemas de IA información adicional para la toma de decisiones.

Por último, es necesario la realización de más estudios económicos, análisis coste-efectividad, en donde se evalúen los costes y la efectividad de los distintos papeles potenciales de los sistemas de IA en la fase del proceso de los programas de detección precoz de cáncer de mama.

III. Análisis de costes

Se ha demostrado que el cribado poblacional de cáncer de mama mediante mamografía mejora el pronóstico y reduce la mortalidad al detectar el cáncer de mama en una fase más temprana y tratable (58). Sin embargo, en estudios retrospectivo (59-61) se ha constatado que alrededor del 20-30 % de los cánceres de intervalo que deberían haberse detectado en las mamografías de cribado no se detectan (FN), y que los hallazgos sospechosos a menudo resultan ser benignos (FP).

Las guías europeas recomiendan la doble lectura de las imágenes mamográficas por dos especialistas en radiología para garantizar una alta sensibilidad del cribado, es decir, para identificar correctamente a las mujeres con cáncer de mama. Esta doble lectura supone una importante carga de trabajo de lectura para los especialistas en radiología, la cual puede ser difícil de sostener debido a la escasez de especialistas en radiología de mama en muchos países.

La IA utilizada para el análisis de imágenes en el cribado mamográfico ha demostrado ser capaz de identificar exámenes que eran normales, es decir, VN. Dado que la gran mayoría de las mujeres que acuden al cribado no tienen cáncer de mama, la adaptación de la lectura única o doble a las puntuaciones de riesgo de la IA podría permitir una lectura de cribado más eficiente (30, 31, 34, 41). La evidencia extraída de estudios retrospectivos sugiere que el cribado mamográfico podría beneficiarse del uso de la IA al reducir la carga de trabajo de la lectura de cribado y el número de cánceres de intervalo y de FP (5, 37, 62, 63). Ahora bien, se necesitan ensayos aleatorizados para evaluar la eficacia del cribado asistido por IA.

En abril de 2021 se llevó a cabo en Suecia un ensayo aleatorizado y controlado (64) con el objetivo de evaluar la seguridad del uso del cribado de mama asistido por IA en comparación con la doble lectura para determinar el efecto sobre la detección del cáncer. En este ensayo se demostró que utilizar un sistema de IA para clasificar los exámenes de cribado en lectura simple o doble como apoyo a la detección en comparación con el cribado mediante lectura doble estándar puede considerarse seguro, al ser la tasa de cáncer detectado en el cribado similar a la obtenida con la lectura doble, sin aumentar las tasas de revaloraciones, FP o reuniones de consenso. Al mismo tiempo, se observó una reducción sustancial de la carga de trabajo de lectura del cribado, lo que podría permitir a los especialistas en radiología liberarse de una no desdeñable cantidad de lecturas.

El sistema de cribado asistido por IA analizado, aunque al menos requiere un especialista en radiología encargado de la detección, potencialmente podría acabar con la necesidad de una doble lectura de la mayoría de las mamografías aliviando la presión sobre la carga de trabajo y permitiendo a los especialistas en radiología centrarse en diagnósticos más avanzados a la vez que se acortan los tiempos de espera de las pacientes.

Como señalan Lang *et al.* (64), dado que los sistemas de IA tienen un coste financiero, se considera preciso determinar la disposición a pagar para reducir la carga de trabajo. El coste-efectividad de los sistemas de IA puede ser determinado una vez haya sido evaluado el coste derivado de la intervención.

III.1. Objetivo

Conocer el coste que se prevé asociado a la realización de un cribado poblacional de cáncer de mama mediante un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble en comparación con la doble lectura estándar.

III.2. Metodología

III.2.1. Perspectiva

El análisis de costes se realiza desde la perspectiva del financiador del SNS.

III.2.2. Horizonte temporal

Se analiza un horizonte temporal a corto plazo, dos años, coincidente con lo que en los programas de cribado de cáncer de mama se considera una vuelta de cribado.

III.2.3. Población

Mujeres candidatas a participar en programas de cribado de cáncer de mama de base poblacional, incluidas las que tienen riesgo hereditario moderado de cáncer de mama o antecedentes de cáncer de mama.

III.2.4. Intervención

Sistema de IA, comercialmente disponible, como apoyo a la detección en el cribado mamográfico. La IA se utiliza para clasificar los exámenes de cribado para lectura única o doble.

Basado en el examen mamográfico, el sistema de IA primero analiza la imagen mamográfica y predice el riesgo de cáncer en una escala continua que va de 1 a 10: entre 1 y 7 se considera que el riesgo de malignidad es bajo, entre 8 y 9 intermedio y para una puntuación de 10 alto. Además, suministra marcas CAD en los hallazgos regionales sospechosos de calcificaciones y de lesiones de tejidos blandos, con una puntuación de riesgo regional en una escala discreta de 1 a 98, para ayudar a los especialistas en radiología a interpretar con precisión las imágenes mamográficas. El sistema de IA se preconfigura para que las marcas CAD solo estén disponibles para los exámenes con puntuaciones de riesgo 8, 9 y 10 con el fin de limitar el número de marcas CAD que puedan perturbar la lectura de pantalla o un aumento de los FP. Los exámenes con el 1 % de riesgo más elevado se marcan en la lista de trabajo de alto riesgo como 10 H. Para seleccionar este grupo se utiliza un umbral de puntuación de riesgo de 9,8, que se determina a partir de la distribución de puntuación de riesgo observada en la población de cribado.

Con base en las puntuaciones de riesgo de malignidad, los exámenes con puntuaciones entre 1 y 9 se someten a lectura única y con puntuación de 10 a lectura doble. La lectura doble la realizan dos especialistas en radiología de mama diferentes, teniendo acceso el segundo especialista en radiología a la evaluación del primero. Los lectores están al corriente de las puntuaciones de riesgo de todos los exámenes recogidas en el PACS y en la imagen del monitor, y primero leen los exámenes sin las marcas CAD y luego con ellas para exámenes con puntuaciones de riesgo 8-10.

Las mujeres son revaloradas para pruebas adicionales en función de los hallazgos sospechosos, teniendo los especialistas en radiología la decisión final de revaloración. Además, los especialistas en radiología revaloran el 1 % de los exámenes con el riesgo más alto, excepto los FP obvios.

Antes de la decisión final, los lectores tienen la opción de remitir los exámenes con hallazgos difíciles o equívocos a una reunión de consenso o a una revaloración técnica, por ejemplo, debido a la mala calidad de la imagen, o a ambas. Dos especialistas en radiología vuelven a evaluar estos exámenes, adoptando una decisión conjunta que implica una nueva revaloración o que se despeje la sospecha de malignidad.

III.2.5. Comparador

Doble lectura, no cegada, de los exámenes de cribado realizada por dos especialistas en radiología sin la ayuda de la IA. Los resultados de la lectura son: no sospecha de malignidad o revaloración. Las participantes pueden ser revaloradas por decisión de los especialistas en radiología debido a hallazgos mamográficos o a síntomas declarados por ellas mismas.

Al igual que para la intervención, los lectores tienen la opción de remitir los exámenes con hallazgos difíciles o equívocos a una reunión de consenso o a una revaloración técnica, o a ambas. Dos especialistas en radiología vuelven a evaluar estos exámenes, adoptando una decisión conjunta que implica una nueva revaloración o que se despeje la sospecha de malignidad.

III.2.6. Técnica

Se utiliza un sistema de mamografía 3D para los exámenes de cribado. El examen de cribado estándar incluye dos proyecciones por mama con la adición de proyecciones de desplazamiento de implantes para las personas con implantes mamarios.

III.2.7. Efectividad

Como medida de efectividad en el análisis de costes se emplea la carga de trabajo que implica la lectura de las imágenes mamográficas de cribado realizada por especialistas en radiología para las estrategias analizadas, entendida como la suma de todas las lecturas de cribado, incluidas las realizadas en las reuniones de consenso. Los datos de efectividad se extraen del ensayo clínico aleatorizado y controlado MASAI, llevado a cabo entre abril de 2021 y 2022 por Lang *et al.* en Suecia (64). De acuerdo con este estudio se observa una reducción del 44,3 % de la carga de trabajo de lectura de pantalla al realizarse en el grupo intervención 36.886 lecturas radiológicas menos que en el grupo control. Aunque no se midió el tiempo real ahorrado debido al uso de la IA, se señala que, si un especialista en radiología leyese una media de 50 exámenes de cribado por hora, éste tardaría entre cuatro y seis meses menos en leer los 46.345 exámenes de cribado del grupo de intervención en comparación con los 83.231 del grupo de control.

Esta reducción de la carga de trabajo de lectura de pantalla ocasionada por la IA se realiza de forma clínicamente segura. La tasa de detección de cáncer para la intervención (6,1 por 1.000 participantes cribadas)

está por encima del límite inferior de seguridad preespecificado en el ensayo (3 por 1.000 participantes cribadas), y es similar a la de la lectura doble sin IA (5,1 por 1.000 participantes cribadas), siendo la diferencia absoluta en la detección de cáncer por 1.000 participantes cribadas de 1,0 (IC al 95 %: 0,0-2,1), diferencia no estadísticamente significativa. Además, el uso de la IA no influye sobre la tasa de revaloraciones, FP o reuniones de consenso dado que: 1) la tasa media de revaloración para el cribado con apoyo de la IA es igual a 2,2 % y a 2,0 % para la doble lectura sin IA, siendo ambas similares a la tasa media de revaloración del 2,1 % observada en la clínica seis meses antes del inicio del ensayo, 2) la tasa de FP es igual a 1,5 % para ambas estrategias y 3) la tasa de reuniones de consenso es del 4,0 % para el cribado con apoyo de IA y del 3,9 % para la doble lectura sin IA. En el ensayo el sistema de IA no provee puntuaciones de riesgo en el 0,8 % de los casos, los cuales fueron analizados mediante lectura doble.

Puede resultar raro que como medida de efectividad no se haya utilizado la razón de cánceres de mama identificados en la exploración de cribado entre cánceres de mama identificados y no identificados en la exploración de cribado, es decir, VP entre VP más FN (cánceres de intervalo). El motivo es que en los resultados publicados en el ensayo MASAI (64), evidencia de mayor calidad publicada hasta la fecha, no se ha recogido el efecto del uso de la IA como herramienta de apoyo a la detección para clasificar los exámenes de cribado en lectura simple o doble sobre el resultado del cribado, es decir, sobre el número de cánceres de intervalo. En este sentido, se señala que la tasa de cánceres de intervalo se evaluará después de que toda la población del estudio de 100.000 participantes sometidos a cribado haya tenido al menos un seguimiento de dos años (estimado en diciembre de 2024).

III.2.8. Costes

De acuerdo con la perspectiva adoptada para el análisis, solo se computan los costes directos sanitarios en los que incurre el SNS para la intervención y el comparador analizados. No se evalúan los costes indirectos sanitarios (los producidos por la morbilidad o mortalidad prematuras ocasionadas por la enfermedad), los costes directos (gastos desplazamiento, de cuidadores, etc.) y los indirectos (pérdida de productividad, coste de oportunidad del tiempo invertido en el tratamiento, etc.) no sanitarios, y los costes intangibles (dolor, ansiedad, etc.).

No se computa el coste de realización de las mamografías basales ni el de las revaloraciones, se asume que se realizan el mismo número de ellas para las estrategias evaluadas. Dicha asunción se hace en función de los re-

sultados del ensayo MASAI (64), descritos en el apartado anterior, en el que se señala no haber diferencias estadísticamente significativas entre la intervención y el comparador analizados con respecto a la tasa de detección, tasa de revaloración, tasa de FP y tasa de reuniones de consenso. Si bien, debiera corroborarse en la práctica habitual ya que las condiciones de ensayo pueden diferir de las condiciones en rutina que pueden resultar menores. Este hecho, sin embargo, es poco plausible ya que las condiciones de los programas de cribado cuentan con sistemas de control de calidad por encima de la práctica habitual.

Para la estrategia de cribado con IA, los costes directos valorados son el coste por estudio mamográfico realizado por la IA y el coste por estudio mamográfico realizado por el especialista en radiología, mientras que para la estrategia mediante doble lectura estándar se computa el coste por estudio mamográfico realizado por el especialista en radiología. El coste por estudio mamográfico realizado por la IA se extrae del expediente de contratación AB-CON3-23-015 de la Consellería de Sanidade de la Xunta de Galicia abierto con el objeto de contratar un software basado en IA para apoyar a los especialistas en radiología del Programa Gallego de Detección Precoz del Cáncer de Mama (PGDPCM) en la lectura de mamografías 3D y/o tomosíntesis, con un alcance de realización de un mínimo de 35.000 estudios mamográficos (65). Este expediente ha sido resuelto a favor de la compañía Bayer Hispania SLU por un importe de 28.798 € con IVA (23.800 € sin IVA). En el expediente de resolución se señala que el sistema de IA adjudicado es Transpara 1.7.4, a través de la herramienta Bayer Calantic, que el coste de instalación del sistema de IA es único para todo el PGDPCM y que está incorporado en el contrato, por lo que no supone coste adicional, y que el coste del mantenimiento del sistema de IA, que incluye soporte de incidencias y actualizaciones de la versión del software durante la duración del contrato, también está incluido en el importe de adjudicación.

El coste por estudio mamográfico realizado por el especialista en radiología se calcula multiplicando el coste del especialista en radiología por el tiempo que emplea en evaluar un estudio mamográfico. El coste del especialista en radiología se calcula en función de su retribución anual y del número de horas anuales trabajadas. La retribución anual del especialista en radiología, igual a 74.757,31 €, se obtiene de la instrucción 1/2023, de 23 de febrero, de la Dirección General de Osakidetza-Servicio Vasco de Salud, sobre retribuciones del personal perteneciente al ente público Osakidetza-Servicio Vasco de Salud para el año 2023. Dicho valor se calcula para la categoría de facultativo especialista en función de su sueldo, complemento de destino, complemento específico del puesto, complemento

de dedicación exclusiva, complemento de productividad y complemento de hospitalización, igual a 56.677,26 €, y de la aportación de la empresa a la Seguridad Social, por contingencias comunes, contingencias AT y EP, mecanismo de equidad intergeneracional (IMEI), cotización por desempleo (tipo general), FOGASA y formación profesional, igual a 18.080,05 € (31,90 % de 56.677,26 €) (aportación extraída del portal del Instituto Nacional de la Seguridad Social (INSS)).

El número de horas anuales trabajadas se calcula para un especialista en radiología con actividad exclusivamente asistencial, en la que dedica a la actividad asistencial directa (realización de pruebas, informes, etc.) seis horas por jornada (86 % de la jornada) y a la actividad asistencial indirecta una hora por jornada (14 % de la jornada) (66). En consecuencia, no se puede imputar directamente todo el coste del especialista en radiología a la realización del estudio mamográfico, por lo que se computa sólo para las horas dedicadas a la actividad asistencial directa. De las 1.592 horas anuales trabajadas (67) el especialista en radiología dedica a la actividad asistencial directa 1.365 horas.

El tiempo que utiliza un especialista en radiología en evaluar un estudio mamográfico, se obtiene del catálogo de exploraciones de la Sociedad Española de Radiología Médica (SERAM), edición 2016 (68). En este se señala que el tiempo médico, definido como el tiempo empleado por el especialista en radiología en realizar el informe radiológico más la supervisión o realización de la exploración en los casos que así lo requiera (69), para una mamografía de cribado poblacional 2P más tomosíntesis es igual a 7 minutos.

En la tabla 15 quedan recogidos los precios y los recursos consumidos para las estrategias analizadas.

Tabla 15. **Precios y recursos consumidos**

		Fuente
Coste sistema IA	**28.798 €**	**Expediente**
Retribución anual especialista en radiología	**74.757,31 €**	
Sueldo	17.049,72 €	Osakidetza
Complemento destino	9.978,64 €	Osakidetza
Complemento específico del puesto	10.460,66 €	Osakidetza
Complemento por dedicación exclusiva	5.114,90 €	Osakidetza
Complemento de productividad	12.369,12 €	Osakidetza
Complemento hospitalización	1.704,22 €	Osakidetza
Aportación a la empresa a la Seguridad Social	18.080,05€	Osakidetza
Horas anuales trabajadas (especialista en radiología con actividad exclusivamente asistencial)	**1.365**	
Jornada trabajo anual (horas)	1.592	Osakidetza
Proporción jornada dedicada por el especialista en radiología a la actividad asistencial directa	0,86	SERAM
Tiempo evaluación estudio mamográfico (mamografía de cribado poblacional 2P más tomosíntesis) (min.)	**7**	**SERAM**

Todos los costes se evalúan en euros del 2023.

III.2.9. Análisis

Se lleva a cabo un análisis económico para estimar el coste incremental que se prevé asociado a la realización de un cribado poblacional de cáncer de mama mediante un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble en comparación con la doble lectura estándar. Dicho análisis, como ha quedado indicado en los apartados anteriores, se realiza desde la perspectiva del financiador del SNS y para un horizonte temporal a corto plazo. No se descuentan los costes debido al horizonte temporal adoptado.

El análisis se ejecuta para una cohorte población de 50.000 mujeres cribadas en el programa de cribado de cáncer de mama de base poblacional. Utilizando la técnica de micro-costes, para la perspectiva y el horizonte temporal señalados se calculan los costes directos sanitarios: el coste por estudio mamográfico realizado por la IA y el coste por estudio mamográ-

fico realizado por el especialista en radiología. La efectividad se mide con base en la carga de trabajo del especialista en radiología asociada a las estrategias analizadas.

Dado que las variables utilizadas para la estimación de la efectividad y de los costes pueden presentar incertidumbre, se realiza un análisis de sensibilidad univariante. En dicho análisis se modifican los valores base adoptados de las variables que presentan mayor incertidumbre: tasa de reducción de la carga de trabajo y coste por estudio mamográfico realizado con el sistema de IA. El análisis de sensibilidad se ejecuta para una tasa de reducción de la carga de trabajo del 29,06 %, obtenida al incrementar para el grupo intervención del ensayo MASAI (64) el número de lecturas totales en un 27,41 % (de 46.345 a 59.047), manteniéndose igual el número de lecturas del grupo comparador (83.231). El incremento señalado se debe a que, de acuerdo con la opinión de expertos, la lectura doble por especialista en radiología en el grupo intervención no sólo se debe realizar para las imágenes mamográficas con riesgo de malignidad alto, puntuación de riesgo de 10, sino también para las que presentan riesgo intermedio, puntuaciones de 8 y 9. Para el coste por estudio mamográfico realizado con el sistema de IA, el valor a computar es de 4,72 £, igual a 4,82 € de 2023 (incluido el coste de mantenimiento), obtenido del estudio de Vargas Palacios *et al.* (70) para el sistema de IA Mia® (Kheiron Medical Technologies).

III.3. Resultados

III.3.1. Efectividad

Para una cohorte poblacional de 50.000 mujeres cribadas, la carga de trabajo de los especialistas en radiología para la estrategia de cribado poblacional de cáncer de mama mediante doble lectura radiológica es de 104.010 lecturas, igual a 100.000 lecturas de cribado, más 2.041 lecturas de revaloración (igual a la tasa de revaloración, 0,2 % por el número de lecturas de cribado), más 1.969 lecturas de consenso (igual a la tasa de reuniones de consenso, 3,9 %, por el número de mujeres cribadas).

Para la estrategia cribado poblacional de cáncer de mama mediante un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble, la carga de trabajo de los especialistas en radiología es de 57.915 lecturas radiológicas, igual al producto de la tasa de reducción de la carga de trabajo de lectura de pantalla, 44,3 % (64), ocasionada al utilizar un sistema de IA como apoyo a la detección frente a la lectura doble, por las 104.100 lecturas realizadas con la do-

ble lectura estándar. Dado que la IA se utiliza para clasificar los exámenes de cribado para lectura única o doble, el número de imágenes mamográficas leídas por la IA es de 50.000, igual a la población cribada.

En la tabla 16 queda recogida la carga de trabajo radiológica para las estrategias de cribado analizadas.

Tabla 16. **Carga de trabajo radiológica para las estrategias de cribado analizadas**

	Intervención	Comparador
Imágenes leídas por especialista en radiología	57.915	104.010
Imágenes leídas por la IA	50.000	—

III.3.2. Costes

El coste por estudio mamográfico realizado con el sistema de IA es de 0,82 €, igual al importe de contratación del expediente AB-CON3-23-015, 28.798 €, entre el número mínimo de lecturas contratadas, 35.000.

El coste por estudio mamográfico realizado por el especialista en radiología es de 6,39 €, igual al coste por minuto del especialista en radiología, 0,91 € (retribución anual del especialista en radiología, 74.757,31 €, entre el número de horas anuales de trabajo dedicados a la actividad asistencial directa por el especialista en radiología con actividad exclusivamente asistencial, 1.365, entre 60 minutos) por el tiempo que tarda el especialista en radiología en evaluar una mamografía de cribado poblacional 2P más tomosíntesis, 7 minutos.

En la tabla 17 queda recogidos los costes directos sanitarios evaluados.

Tabla 17. **Costes directos sanitarios**

	Costes directos sanitarios (€ de 2023)
Coste por estudio mamográfico realizado por el sistema de IA	0,82 €
Coste por estudio mamográfico realizado por especialista en radiología	6,39 €

El coste de la intervención analizada para la cohorte poblacional de 50.000 mujeres cribadas es de 411.188,66 €, igual al coste de los estudios

mamográficos realizados por la IA, 41.140 € (coste por estudio mamográfico realizado por el sistema de IA, 0,82 €, por el número de imágenes leídas por la IA, 50.000), más el coste de los estudios mamográficos realizados por especialista en radiología, 370.048,66 € (coste por estudio mamográfico realizado por especialista en radiología, 6,39 €, por el número de imágenes leídas por especialista en radiología, 57.915).

Para la misma cohorte poblacional de mujeres cribadas, el coste del comparador, coste de los estudios mamográficos realizados por especialista en radiología, es de 664.573,28 €, igual al coste por estudio mamográfico realizado especialista en radiología, 6,39 €, por el número de imágenes leídas por especialistas en radiología, 104.010.

En la tabla 18 quedan reflejados los costes totales para las intervenciones analizadas.

Tabla 18. **Costes totales para las intervenciones analizadas**

	Intervención (€ de 2023)	Comparador (€ de 2023)
Coste estudios mamográficos realizados por la IA	41.140,00 €	
Coste estudios mamográficos realizados por especialista en radiología	370.048,66 €	664.573,28 €
TOTAL	**411.188,66 €**	**664.573,28 €**

III.3.3. Análisis de costes

El resultado del análisis realizado indica que el coste incremental para la estrategia cribado poblacional de cáncer de mama mediante un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble frente a la estrategia cribado poblacional de cáncer de mama mediante doble lectura radiológica, es de –253.384,62 € (ver tabla 19).

Tabla 19. **Análisis de costes, coste incremental**

	(€ de 2023)
Coste intervención	418.372,53 €
Coste comparador	664.573,28 €
Coste incremental	–253.384,62 €

III.3.4. Análisis de sensibilidad

El análisis de sensibilidad indica que cuando el coste por estudio mamográfico realizado por el sistema de IA es igual a 4,82 €, coste indicado para el sistema de IA MIA® en el estudio de Vargas-Palacios *et al.* (70), el coste incremental aumenta un 78,88 %, pasando de –253.384,62 € a –53.524,62 €. Cuando la tasa de reducción de la carga de trabajo es igual a 27,41 %, al computar en el grupo intervención para la doble lectura además de los exámenes de cribado con riesgo de malignidad alto (puntuación de riesgo 10), también los de riesgo de malignidad intermedio (puntuaciones 8 y 9), el número de lecturas realizadas por especialista en radiología aumenta un 30,37 % (de 57.915 a 75.501) y en consecuencia el coste incremental aumenta un 44,35 %, pasando de –253.384,62 € a –141.018,64 € (ver tabla 20).

Tabla 20. **Análisis de sensibilidad univariante**

	Caso base		Tasa reducción carga trabajo = 27,41 %		Coste por estudio mamográfico realizado por el sistema de IA = 4,82 €	
	Intervención (€ de 2023)	Comparador (€ de 2023)	Intervención (€ de 2023)	Comparador (€ de 2023)	Intervención (€ de 2023)	Comparador (€ de 2023)
Coste estudios mamográficos realizados por la IA	41.140,00 €		41.140,00 €		241.000,00 €	
Coste estudios mamográficos realizados por especialista en radiología	370.048,66 €	664.573,28 €	482.414,65 €	664.573,28 €	370.048,66 €	664.573,28 €
TOTAL	**411.188,66 €**	**664.573,28 €**	**523.554,65 €**	**664.573,28 €**	**611.048,66 €**	**664.573,28 €**
Coste incremental	–253.384,62 €		–141.018,64 €		–53.524,62 €	

III.4. Discusión

En este estudio se evalúa el coste incremental asociado a la realización de un cribado poblacional de cáncer de mama mediante una estrategia de cribado que emplea un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura única o doble en comparación con la estrategia de doble lectura estándar. La evalua-

ción llevada a cabo obtiene como resultado para la cohorte analizada de 50.000 mujeres cribadas, un coste incremental igual a –253.384,62 €, es decir, el gasto sanitario de la estrategia de cribado mediante un sistema de IA es 253.384,62 € menor que el de la estrategia de doble lectura estándar. Los resultados obtenidos muestran que el gasto ocasionado por la IA de 41.140 €, igual al coste de los estudios mamográficos realizados por la IA, queda compensado por el gasto evitado, –294.524,62 €, al verse reducida la carga de trabajo de lecturas de imágenes mamográficas realizadas por especialista en radiología en 46.095 en el grupo intervención frente al grupo control.

En la evaluación realizada se analiza cómo varía el coste incremental calculado si la tasa de reducción de la carga de trabajo tomada como valor base en el estudio, 44,3 %, pasa a ser del 27,41 % al contabilizarse para doble lectura en el grupo intervención las imágenes mamográficas con riesgo de malignidad intermedio (puntuación de riesgo del sistema de IA 8 y 9), manteniéndose el resto de las variables utilizadas para el cálculo de costes constantes. El resultado obtenido muestra que, aunque el número de exámenes mamográficos realizados por especialistas en radiología en el grupo intervención aumenta un 30,37 %, el gasto sanitario ocasionado por la intervención sigue siendo inferior frente al comparador, –141.018,64 €. En el análisis, esta disminución en el gasto sanitario se produce mientras la tasa de reducción de la carga de trabajo sea superior al 6,2 %. Para una tasa de reducción de la carga de trabajo inferior a esta, el gasto sanitario es menor para la estrategia de doble lectura estándar en comparación con la estrategia de cribado con IA.

Además, se observa que el coste de la estrategia de cribado de cáncer de mama mediante un sistema de IA como apoyo a la detección para clasificar los exámenes de cribado para lectura única o doble lectura depende principalmente del coste por examen mamográfico leído por la IA. Cuando este coste es igual a 4,82 €, el coste incremental aumenta un 78,88 %, siendo igual a –53.524,62 €. Esta disminución en el gasto sanitario se produce siempre que el coste por examen mamográfico leído por la IA sea inferior a 5,89 €. Para un coste superior a éste, el gasto sanitario es menor para la estrategia de doble lectura estándar en comparación con la estrategia de cribado con IA.

Como se aprecia, el coste por examen mamográfico realizado con el sistema de IA presenta bastante incertidumbre, debido a que según parece este depende en gran medida de cómo se negocie con las empresas proveedoras. En el análisis este coste varía desde los 0,82 €, obtenido de un expediente de contratación licitado por la Consellería de Sanidade de la Xunta de Galicia, hasta los 4,82 %, extraído del estudio de Vargas-Pa-

lacios *et al.* (70) en el que se señala que dicho coste se obtuvo mediante la negoción de una estructura de precios con el proveedor del sistema de IA. En el análisis base realizado en el coste por examen mamográfico realizado por la IA, tanto el coste de la instalación del sistema de IA como su mantenimiento está incluido en el mismo. Un punto a tener en cuenta cuando se negocie este coste es si en el mismo están incluidos o no ambos. En el estudio de Vargas-Palacios *et al.* (70) se señala la influencia que estos dos costes pueden tener en su análisis. Indican que el coste de la instalación apenas modifica los resultados obtenidos, mientras que un coste anual de mantenimiento de 17.000 £ anuales (17.446,53 € de 2023) ocasiona que la doble lectura sea coste-efectiva en comparación con la estrategia de IA analizada en él.

Dado que los sistemas de IA pueden jugar un papel potencial en distintas fases del proceso de cribado de mama, cómo se propone utilizar los sistemas de IA y cómo interactúan los especialistas en radiología con estos sistemas son aspectos a tener en cuenta en la evaluación, ya que la combinación de expertos e IA en la práctica clínica es lo que determinará la precisión global y los resultados para las mujeres. En este estudio se analiza una de las posibles estrategias para integrar la IA en el proceso del cribado de mama. El diseño de la intervención, tal y como se ha descrito en la metodología de este análisis de costes, permite hacer frente a alguno de los problemas éticos con los que se suele topar el uso de la IA. Por ejemplo, el uso rutinario del aprendizaje automático puede cambiar las habilidades humanas, ocasionando que los profesionales sanitarios pierdan habilidades que no utilizan regularmente o el denominado sesgo de automatización por el que los humanos tienden a aceptar las decisiones de las máquinas, incluso si son erróneas. En el ensayo MASAI (64) se señala que el acceso por parte de los especialistas en radiología a las puntuaciones de riesgo y a las marcas CAD no parece introducir un sesgo de automatización perjudicial, ya que la tasa de FP permanece inalterada. Indican que esto subraya la importancia de que los especialistas en radiología tengan la decisión final de revaloración, ya que además de reducir los FP, constituye un enfoque práctico para cumplir con los requisitos médicos-legales establecidos, frente a las actuales incertidumbres éticas y legales del uso de la IA como lector autónomo. Que la estrategia de cribado haga hincapié en el papel central del especialista en radiología para tomar la decisión final de revalorar o no a una paciente y que los resultados del cribado dependan de la actuación de los especialistas en radiología participantes y no de la IA implica que el uso de la IA no ponga en tela de juicio el modelo tradicional de responsabilidad médica, en el cual las personas que reciben la intervención son pacientes de un profesional sanitario identificado que se responsabiliza de ese aspecto de su atención.

Otra de las valoraciones a realizar, en casos en los que los costes de la IA por mamografía fueran más elevados e incluso considerados no coste-efectivos, es la posibilidad de tomar una decisión por la escasez de profesionales de radiología o la necesidad de emplearlos en otras indicaciones no ligadas a los programas de cribado como opciones de mayor valor añadido, en este sentido, podría justificarse una decisión positiva, sin embargo, no sería la deseable como tal.

Una limitación del análisis de costes es que la medida de efectividad utilizada para su realización es la carga de trabajo radiológica ocasionada por el número de lecturas de imágenes mamográficas y no la capacidad que la estrategia de cribado de cáncer mediante la IA tiene de identificar un cáncer durante su fase detectable. La evaluación de la tasa de cánceres de intervalo, junto con una caracterización de los cánceres detectados en toda la población del estudio, proporciona más información sobre la eficacia del cribado, los posibles efectos secundarios, como el sobrediagnóstico, y las implicaciones pronósticas del uso de la IA en el cribado mamográfico.

Que sólo se haya analizado el efecto de los sistemas de IA sobre la carga de trabajo ha implicado que solo se hayan analizado los costes de lectura de imágenes mamográficas de la IA y de los especialistas en radiología. Como consecuencia de esto no se han medido los costes ocasionados por el diagnóstico, tratamiento y seguimiento hospitalario ocasionados por los cánceres. Además, no se ha computado el coste de realización de las mamografías basales, ni el de las mamografías radiológicas y ecografías adicionales debido a las revaloraciones, al asumirse para las estrategias analizadas que la eficacia del cribado es similar, al no encontrarse en el ensayo en el que se basa este estudio diferencias estadísticamente significativas con respecto a la tasa de detección, tasa de revaloración, tasa de FP y tasa de reuniones de consenso.

Por otro lado, cabe señalar también como una limitación que los datos de efectividad se extraen de un único ensayo, que como tal presenta sus propias limitaciones. Así, el ensayo en el que se basa el estudio se realiza en un único centro, para un único dispositivo mamográfico y sistema de IA único, en el cual los especialistas en radiología participantes tienen una experiencia en la lectura de imágenes mamográficas entre moderada y alta, y en el que no se recoge información sobre la raza y etnia de las mujeres participantes. Todo esto puede limitar la generalización de los resultados obtenidos en él a otras realidades sanitarias.

Además, hay que indicar que el análisis de costes sólo se realiza para la estrategia en la que la IA se utiliza como apoyo a la detección en el cribado mamográfico para clasificar los exámenes de cribado para lectura

única o doble. Esto puede suponer un sesgo al no haber incluido en el mismo el análisis de otras estrategias que, de acuerdo con la literatura revisada en el apartado de revisión de la evidencia científica de este informe, también disminuyen la carga de trabajo de los especialistas en radiología, como sucede cuando la IA se utiliza como herramienta de clasificación previa al cribado, en el que la IA se aplica para eliminar todos los casos normales evidentes, es decir los de bajo riesgo (31, 43). Que el análisis de costes se haya realizado para la estrategia señalada se debió a que para el mismo nos basamos en la evidencia científica de mayor calidad metodológica encontrada hasta la fecha, proporcionada por el ensayo MASAI (64) el cual sólo analiza la estrategia evaluada, descartando el resto de los estudios, todos ellos de cohortes retrospectivos, y por tanto de menor calidad metodológica. Aunque no se haya analizado el coste para la estrategia en la que la IA se utiliza para la clasificación previa al cribado, dado que para esta la reducción en la carga de trabajo observada en los estudios analizados (entre el 62 % y el 71 %) (31, 43) es superior a la obtenida en el ensayo MASAI (44,3 %), se puede suponer que para ella los resultados obtenidos en el análisis de costes se mantendrían e incluso serían mejores.

Por último, y dado que, en las Comunidades y Ciudades Autónomas, no existe un único modelo de cribado de cáncer, al variar entre frecuencias anuales y bienales, en el rango de edades de cribado y también en la técnica, puede que sea cuestionable en algunos entornos la validez externa de algunas de las conclusiones del análisis de costes realizado, principalmente en aquellos en los que se emplee lectura radiológica única.

III.5. Conclusiones

Para el caso base analizado, el coste asociado a la realización de un cribado poblacional de cáncer de mama mediante una estrategia de cribado que emplea un sistema de IA como apoyo a la detección utilizado para clasificar los exámenes de cribado para lectura por especialista en radiología única o doble es menor en comparación con la estrategia de doble lectura estándar.

El coste ocasionado por el sistema de IA, 41.140 €, queda compensado por el menor coste de lectura de imágenes mamográficas realizada por especialista en radiología en el grupo intervención, consecuencia de realizarse 46.095 lecturas menos en este en comparación con el grupo control.

IV. Implicaciones éticas

La IA y su uso es un tema de debate en la actualidad. El desarrollo y empleo de la IA progresa de manera exponencial en la provisión de servicios en salud y más concretamente en el de la radiología en el que la escasez de profesionales y la importancia de contar con un valor añadido cuando actúan es crucial. Un área de la radiología en el que la IA está cobrando una gran importancia es el de la atención del cáncer de mama, especialmente en lo que respecta a su detección y cribado. La aplicación de los sistemas de IA en este campo está creando una serie de expectativas en cuanto a un mejor rendimiento analítico, beneficios para los pacientes, los profesionales sanitarios y la sociedad, y un cambio en el papel profesional de las y los especialistas en radiología. Si bien esta tecnología tiene el potencial de mejorar los servicios en salud en su eficiencia, los administradores de los sistemas sanitarios, los profesionales sanitarios y los desarrolladores de los algoritmos de IA deben reconocer que la atención al cáncer de mama mediante IA es un reto con implicaciones éticas, legales y sociales, no solo técnicas, que deben considerarse cuidadosamente para evitar consecuencias perjudiciales para los individuos y grupos, especialmente para los menos favorecidos.

Como paso previo al análisis de las implicaciones éticas, legales y sociales de la implementación de la IA en la práctica clínica, Carter *et al.* (71) mencionan dos aspectos que ayudan a enfocarlo: definir qué es la IA y señalar cómo se desarrolla y utiliza para la atención del cáncer de mama. A grandes rasgos la IA se caracteriza por el uso de técnicas de aprendizaje automático, especialmente aprendizaje profundo, que permite a un algoritmo de forma independiente clasificar y agrupar datos. Estos algoritmos pueden reconocer patrones de estos datos y con el tiempo aprender a identificar y extraer atributos relevantes de los datos para alcanzar el objetivo. El aprendizaje, no supervisado inicialmente por humanos, permite al algoritmo agrupar datos basados en similitudes y reducir, de forma independiente, la dimensionalidad de los datos utilizados para informar una decisión. Los resultados obtenidos de estos sistemas de IA, especialmente las redes neuronales, pueden ser difíciles de interpretar, por su complejidad creciente y por su capacidad para el aprendizaje no supervisado y no restringido. A estos algoritmos se les denomina «cajas negras» porque el detalle de sus operaciones y cómo se relacionan con los resultados no es siempre entendible por el ser humano.

El uso de la IA en el caso del cáncer de mama se dirige especialmente hacia el cribado y diagnóstico, aunque también se emplea para el cálculo

de riesgo, pronóstico y apoyo a la decisión clínica, planificación de la gestión. Se están desarrollando productos basados en la IA para el proceso de imágenes incluida la lectura mamográfica y el diagnóstico de tejido patológico. Hay sistemas de IA para la detección de cáncer de mama disponibles para la aplicación clínica, y el márquetin de las compañías señala que estos productos pueden aumentar la eficiencia, la precisión e ingresos, aunque la evidencia encontrada hasta la fecha no ofrece las mismas conclusiones. Además, se están desarrollando sistemas de IA para mejorar los flujos de trabajo y para ofrecer servicios basados en la IA utilizando tecnologías de pruebas novedosas y no probadas que se comercializan directamente a los consumidores.

A continuación, se examinan cuestiones como los problemas implícitos a la naturaleza de «caja negra» de los algoritmos de IA, el potencial de sesgo, las cuestiones legales, la responsabilidad profesional y la rendición de cuentas, la confianza de los pacientes y profesionales sanitarios en la IA y las implicaciones sobre la justicia y equidad de los sistemas de IA.

IV.1. Caja negra

Como se señala en el estudio de Lokaj B *et al.* (72), la naturaleza de «cajas negras» de los modernos algoritmos de IA es la mayor barrera para su implantación clínica debido a la falta de explicaciones sobre las decisiones que toman y la posible presencia de sesgos, lo que erosiona la confianza de los profesionales sanitarios y pacientes. Goisauf *et al.* (73) apuntan que el problema de las cajas negras en medicina se enmarca aplicando el concepto de opacidad, el cual se puede diferenciar como falta de divulgación, opacidad epistémica y opacidad explicativa. Definen la falta de divulgación como la falta de transparencia en relación con el uso de datos, siendo la privacidad de los pacientes y el conocimiento del uso de sus datos, su consentimiento y la propiedad de estos, preocupaciones asociadas. Para garantizar el uso ético de los datos los pacientes deben saber quién tiene acceso a ellos y si están anonimizados, para lo que se necesitan normas de seguridad, protección de la intimidad y uso ético de la información sensible. La opacidad epistémica es la falta de compresión de cómo funciona el sistema de IA y está causada por la oscuridad procedimental, es decir, las reglas que sigue el sistema de IA no están disponibles, o por la ignorancia procedimental, es decir, las reglas están disponibles, pero no se pueden entender o su entendimiento es complejo para una persona no experta. Por último, la opacidad explicativa la definen como la falta de explicación clínica, es decir, un sistema puede encontrar patrones que no tienen una explicación

clínica con los conocimientos sanitarios actuales. En resumen, el problema de la caja negra, como indican Carter *et al.* (71), está relacionado con que los sistemas de IA puedan ser adecuadamente interpretados y explicados. Como señalan Goisauf *et al.* (73) a menudo estos conceptos son utilizados simultáneamente. Sin embargo, la interpretabilidad se refiere a lo bien que se puede entender cómo funciona un sistema de IA mientras que la explicación se refiere a lo bien que se puede explicar lo que ocurre en la toma de decisiones de la IA en términos comprensibles. La falta de comprensión y transparencia sobre el modo en que un sistema de IA toma una decisión plantea un importante dilema ético.

Carter *et al.* (71) y Goisauf *et al.* (73) indican que en la actualidad los algoritmos de IA menos explicables son los que parecen ser más precisos, por lo que no está claro si la explicación y la precisión deben estar inevitablemente intercambiados o si es posible obtener ambos. Como señalan Carter *et al.* (71), las explicaciones de los profesionales sanitarios sobre sus decisiones no son perfectas, pero ellos son responsables legal, ética y profesionalmente de ellas, son capaces de dar una explicación y pueden ser requeridos para hacerlo. En cambio, los sistemas de IA pueden recomendar decisiones individualizadas de diagnóstico, pronóstico o gestión que no puedan explicarse. Goisauf *et al.* (73) advierten que esto puede plantear cierto riesgo para las y los profesionales de radiología, de quienes se espera que validen algo que no pueden entender, dando por supuesto que estos tienen capacidad de entender la tecnología, sus beneficios y riesgos potenciales. Añaden que esto también se asocia con privar a los pacientes de la capacidad de tomar decisiones basadas en información y justificaciones suficientes, lo cual contradice el requisito ético de que los pacientes ejerzan su autonomía dentro del consentimiento informado.

IV.2. Sesgos

Lokaj *et al.* (72) indican que los sistemas de IA son susceptibles de sesgos, particularmente en casos donde los datos están desequilibrados o ciertos segmentos de la población están infrarrepresentados. Goisauf *et al.* (73) señalan que los sesgos se encuentran asociados a la falta de transparencia de los modelos de IA al ser considerados como «cajas negras». Además, señalan que los sistemas de IA radiológicos también pueden estar sesgados por atributos clínicos de confusión (comorbilidades) y por factores técnicos debido a diferencias sutiles en los datos brutos y post-procesados que surgen del uso de diferentes técnicas de exploración. Mencionan que se ha demostrado que muchos algoritmos sanitarios codifican, refuer-

zan e incluso agravan las desigualdades del sistema sanitario y que pueden empeorar los resultados de los pacientes vulnerables. Estos sesgos se introducen debido a los datos utilizados en el entrenamiento del algoritmo y a las etiquetas dadas a esos datos, que pueden estar cargados con valores, preferencias y creencias humanas y por lo tanto los resultados generados acabarán reflejando las estructuras sociales y políticas, incluidas las injusticias y desigualdades. Consideran que los sistemas de IA no pueden proveer resultados completamente no sesgados u objetivos a partir de **datos incompletos y no representativos**, sino que reflejan los sesgos humanos implícitos en la toma de decisiones.

El sesgo del sistema de IA surge cuando los modelos son seleccionados para representar mejor a la mayoría, pero no a los grupos no representados, los datos son insuficientes para minorías y por la interferencia entre las variables no observadas y las predicciones del modelo. Los sistemas de IA están a menudo desarrollados por compañías en países occidentales y testados con datos de individuos caucásicos lo que provoca desequilibrios de representación en los conjuntos de datos. Como indican Carter *et al.* (71) un algoritmo entrenado y probado en un contexto concreto no es aplicable directamente a otro. Para que un algoritmo de IA sea transferible este requerirá una formación deliberada sobre los datos de la nueva cohorte y el entorno. Los sistemas de IA requieren un cuidadoso desarrollo, prueba y evaluación en cada nuevo contexto antes de su uso en la atención al paciente individual.

Por último, Goisauf *et al.* (73) señalan que los sistemas de IA no pueden corregir los sesgos de igualdad y equidad por sí solos, pero que sí pueden hacerlo las y los investigadores que desarrollan dichos sistemas y las empresas que los comercializan. Se debe garantizar la diversidad a la hora de recopilar datos y abordar los sesgos en el diseño, la validación y el despliegue de los sistemas de IA. Los algoritmos de IA se deben diseñar teniendo en cuenta la comunidad global, y la validación clínica debe hacerse utilizando una **población representativa** de la prevista para el despliegue.

IV.3. Aspectos legales

En abril de 2021 la Comisión Europea propuso el primer marco regulador de la Unión Europea (UE) para la IA, bajo el nombre «Propuesta de reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión»,

2021/0106 (COD). El marco reglamentario sobre IA se propone con los siguientes objetivos: garantizar que los sistemas de IA introducidos y usados en el mercado de la UE sean seguros y respeten la legislación vigente en materia de derechos fundamentales y valores de la Unión, garantizar la seguridad jurídica para facilitar la inversión e innovación en IA, mejorar la gobernanza y la aplicación efectiva de la legislación vigente en materia de derechos fundamentales y los requisitos de seguridad aplicable a los sistemas de IA, y facilitar el desarrollo de un mercado único para hacer uso legal, seguro y fiable de las aplicaciones de IA y evitar la fragmentación del mercado.

La propuesta presentaba un enfoque normativo que se limitaba a establecer los requisitos mínimos necesarios para subsanar los riesgos y problemas vinculados a la IA y establecía normas armonizadas para el desarrollo, la introducción en el mercado y la utilización de sistemas de IA en la Unión a partir de un enfoque proporcionado basado en riesgos. Únicamente imponían cargas normativas cuando era probable que un sistema de IA entrañase altos riesgos al objeto de garantizar un nivel elevado y coherente de protección de los intereses públicos en lo que respectaba a la salud, la seguridad y los derechos fundamentales.

La norma establecía cuatro niveles de riesgo, que marcaban los diferentes niveles de peligrosidad y, en consecuencia, de permisibilidad. Estos niveles eran: riesgo inaceptable, alto riesgo, riesgo limitado y riesgo mínimo. La UE marca obligaciones para los proveedores y personas usuarias en función del riesgo que se acarree.

Un sistema de IA se consideraba de alto riesgo cuando: a) estaba destinado a ser utilizado como componente de seguridad de uno de los productos contemplados en la legislación de armonización de la Unión que se indicaba en el anexo II de la misma, o es en sí mismo uno de dichos productos, y b) conforme a la legislación de armonización de la Unión, indicada en el anexo II, el producto del que el sistema de IA fuese componente de seguridad, o el propio sistema de IA como producto, debía someterse a un evaluación de la conformidad realizada por un organismo independiente para su introducción en el mercado o puesta en servicio. En el anexo II señalado, entre otros, se recogen los dispositivos sanitarios, los cuales están definidos de acuerdo con el marco regulador de la Unión para productos sanitarios, Reglamento (UE) 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios.

En esta propuesta de reglamento se indicaba que con el objeto de mitigar los riesgos que presentaban para los usuarios y las personas afectadas los sistemas de IA de alto riesgo que se introducían en el mercado o ponían en servicio en la Unión, era preciso aplicar ciertos requisitos referentes a la

calidad de los conjuntos de datos utilizados, la documentación técnica y el registro, la transparencia y la comunicación de información a los usuarios, la vigilancia humana, la solidez, la precisión y la ciberseguridad. Además, se consideraba que antes de su introducción en el mercado o puesta en servicio, los sistemas de IA de alto riesgo debían someterse a una evaluación de conformidad que garantizase que eran altamente fiables.

El viernes, 8 de diciembre de 2023, las instituciones de la UE pactaron el texto de la Ley de inteligencia artificial. Esta ley se sustenta en el reglamento propuesto por la Comisión Europea el 21 de abril de 2021 y que establece los límites legales y éticos de la IA. La norma pretende garantizar que la IA desarrollada y utilizada en Europa esté en consonancia plena con los derechos y valores de la UE. La Ley de IA de la UE tiene como objetivo establecer un marco regulador global y con visión de futuro, de principios éticos y obligaciones jurídicas para el desarrollo, el despliegue y el uso de IA, la robótica y las tecnologías conexas en la Unión. La ética de la IA es un conjunto de directrices que asesora sobre el diseño y los resultados de la IA. La UE pretende regular la IA para garantizar mejores condiciones de desarrollo y uso de esta tecnología. La prioridad es garantizar que los sistemas sean seguros, transparentes y trazables. Así, la ley marca unas barreras, delimitadas por los riesgos que la IA puede conllevar para los derechos fundamentales de las personas, incluidas la seguridad y la salud.

En cuanto a la gestión de riesgos para los derechos y libertades de los interesados aplicable a cualquier tratamiento, independientemente de su nivel de riesgo, la Agencia Española de Protección de Datos (AEPD) ha desarrollado una guía. Además, para los casos de tratamientos de alto riesgo, incorpora las orientaciones necesarias para realizar la evaluación de impacto para la protección de datos y, en su caso, la consulta previa a la que se refiere el artículo 36 del Reglamento General de Protección de Datos.

Finalmente, otro punto controvertido es la certificación de sistemas de IA que es un proceso mediante el cual una entidad independiente evalúa y verifica que un sistema cumple con ciertos estándares y regulaciones. Actualmente, no existe un sistema único y global para la certificación de sistemas de IA, pero hay varias iniciativas en curso para desarrollar estándares y marcos para la certificación, como explica la AEPD. Por ejemplo, el Instituto Europeo de Normas de Telecomunicación está trabajando en el desarrollo de normas para la certificación de sistemas de IA.

El agente encargado de velar por el cumplimiento de la norma será la Agencia Española de Supervisión de la Inteligencia Artificial creada por el

Gobierno español el 28 de diciembre de 2021, y adscrita a la Secretaría de Estado de Digitalización e Inteligencia Artificial dentro del Ministerio de Asuntos Económicos y Transformación Digital.

La ley también establece un sistema de sanciones para aquellos que incumplan sus disposiciones. Las sanciones pueden incluir multas y otras medidas administrativas, dependiendo de la gravedad del incumplimiento.

Las multas pueden ser considerables, y llegar hasta el 6 % de los ingresos globales anuales de una empresa o 30 millones de euros, lo que sea mayor.

El acuerdo provisional establece que la ley sobre IA debería aplicarse dos años después de su entrada en vigor, con algunas excepciones para disposiciones específicas. Previsiblemente la norma estará en pleno vigor en 2026, pero tiene que ser ratificada por el Parlamento y los estados miembros.

IV.4. Responsabilidad profesional y rendición de cuentas

Carter *et al.* (71) señalan que la IA potencialmente puede poner en tela de juicio los conceptos tradicionales de la responsabilidad médica. Indican que si la IA puede de forma fiable y coste-efectiva realizar el cribado de mamografía, puede ser concebible que la mamografía de la persona solo sea leída por la IA y que los resultados negativos sean comunicados directamente a ella. Esto puede crear grupos de personas a las que se realizan pruebas médicas independientemente de cualquier lector humano, socavando el modelo tradicional en el cual las personas que reciben la intervención son pacientes de un profesional sanitario identificado que se responsabiliza de ese aspecto de su atención. Especulan que, si los profesionales sanitarios están directamente involucrados en el cuidado del paciente, pero las decisiones dependen de las recomendaciones de una IA no explicable, estos se pueden enfrentar a retos relacionados con la responsabilidad moral y legal. Se puede esperar que asuman responsabilidades que no pueden controlar o explicar, y si no las asumen no quedará claro en quien se debe delegar la responsabilidad. Un cambio en las atribuciones de responsabilidad puede afectar a la confianza de los pacientes en los profesionales sanitarios y proveedores de servicios sanitarios y cambiar los roles profesionales.

Además, indican que la IA tiene implicaciones para las capacidades humanas. El uso rutinario del aprendizaje automático puede cambiar las

habilidades humanas, ocasionando que los profesionales sanitarios pierdan habilidades que no utilizan regularmente o el denominado sesgo de automatización por el que los humanos tienden a aceptar las decisiones de las máquinas, incluso si son erróneas. Por otro lado, el diseño del flujo de trabajo utilizando la IA para clasificar las mamografías normales puede significar que los lectores humanos vean proporcionalmente más cáncer de mama y mejoren en la identificación del cáncer, pero estén menos familiarizados con las variantes de las imágenes normales. La paradoja de la decesión de competencias puede hacer que las mismas en momentos puntuales en los que la alternativa no funcione (p. ej. que la IA artificial no actúe por problemas tecnológicos), no ofrezcan soluciones que hagan que la asistencia sea posible o que se dilate en el tiempo la decisión, en procesos en los que el tiempo es crítico.

IV.5. Confianza

Como señalan Goisauf *et al.* (73), la confianza es un requisito clave para el uso ético de la IA. Las «cajas negras» y la falta de interpretabilidad y explicación pueden provocar una falta de confianza en los sistemas de IA y su aceptación por parte de profesionales sanitarios y pacientes. Para abordar estas barreras, Lokaj *et al.* (72) indican que es crucial centrarse en algoritmos de IA transparentes con robustos métodos explicativos y sistemas de interfaz de IA fáciles de usar durante la fase de concepción, además de poner una mayor atención en la educación de los profesionales sanitarios y de los pacientes sobre la IA. Goisauf *et al.* (73) dicen que la IA implica un elemento de incertidumbre y riesgo para el paciente vulnerable por lo que consideran que la explicación es clave para fomentar la confianza en un sistema de IA. La falta de explicación, de transparencia y de comprensión humana del funcionamiento de los sistemas de IA provocará inevitablemente que los profesionales sanitarios no confíen en las decisiones tomadas por la IA, ni en la fiabilidad y precisión de dichos sistemas.

Por último, Carter *et al.* (71) señalan que parece probable que para mantener la confianza en los sistemas sanitarios se requiera al menos cierta responsabilidad pública sobre el uso de la IA en dichos sistemas. Los profesionales sanitarios y los pacientes debieran estar informados en todo momento de los hallazgos y de dónde parten cada uno de los resultados finales desde los cuales se sustentan las decisiones y, además, por su consecuencia legal, debieran estar plasmados en los consentimientos informados.

IV.6. Justicia y equidad

La justicia es uno de los cuatro principios de la bioética: autonomía, beneficencia, no maleficencia y justicia. La justicia se refiere a la idea que los beneficios y costes de la investigación y de los cuidados médicos deben ser distribuidos equitativamente, incluso trasciende de la equidad, ya que su fin último es superar las barreras que sustentan la diferencia.

El principio de justicia aparece a menudo asociado con beneficencia y no maleficencia, ya que la injusta distribución de recursos conduce a la discriminación y puede causar daño. La asociación entre injusticia, discriminación y decisiones injustas tomadas por los sistemas de IA también se ha relacionado con los sesgos. Los sistemas de IA sesgados pueden conducir a actuaciones injustas, discriminatorias o a decisiones erróneas.

La integración de los sistemas de IA en la atención sanitaria incurre en el riesgo de replicar discriminaciones que ya existen en la sociedad, por lo que el desarrollo de la IA tiene que promover la justicia mientras elimina discriminaciones injustas, garantizando beneficios compartidos y previniendo infringir nuevos daños que pueden surgir de los sesgos implícitos. Las herramientas de IA pueden decidir en favor de un grupo de pacientes debido a sesgos implícitos en lugar de priorizar una emergencia real en radiología, reflejando la necesidad de que todos los implicados en el proceso se adhieran a directrices éticas que promuevan la justicia. Es por ello que la adaptabilidad a los contextos cambiantes deber ser un principio a considerar en las contrataciones públicas de soluciones tecnológicas basadas en IA, de cara a no incrementar o ahondar en la discriminación derivada de una disfunción y distopia no suficientemente contextualizada.

V. Referencias bibliográficas

(1) Sociedad Española de Oncología Médica (SEOM). Las cifras del cáncer en España 2023. Disponible en: https://seom.org/images/Las_cifras_del_Cancer_en_Espana_2023.pdf; 2023. [Consultada 26/05/2023].

(2) Red de programas de Cribado de Cáncer. Programas de Cribado de Cáncer de Mama. Informe de evaluación 2017 (datos: abril 2018). Disponible en: https://cribadocancer.es.

(3) O. Díaz, A. Rodríguez-Ruiz, A. Gubern-Mérida, R. Martí, M. Chevalier. ¿Son los sistemas de inteligencia artificial una herramienta útil para los programas de cribado de cáncer de mama? Radiología. 2021, 63 (3): 236-244. doi: 10.1016/j.rx.2020.11.006.

(4) Fez Herráiz Julia de, Rodríguez Alcalá Francisco Javier. El cribado de cáncer de mama a examen. Rev Clin Med Fam [Internet]. 2019; 12(3): 115-118. Disponible en: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1699-695X2019000300115&lng=es.

(5) Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, *et al.* Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. Bmj. 2021 Sep 1;374:n1872. doi: 10.1136/bmj.n1872.

(6) Taylor-Phillips S, Seedat F, Kijauskaite G, Marshall J, Halligan S, Hyde C, *et al.* UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. Lancet Digit Health. 2022 Jul;4(7):e558-e565. doi: 10.1016/S2589-7500(22)00088-7.

(7) Jairam MP, Ha R. A review of artificial intelligence in mammography. Clin Imaging. 2022 Aug;88:36-44. doi: 10.1016/j.clinimag.2022.05.005.

(8) Hickman SE, Woitek R, Le EPV, Im YR, Mouritsen Luxhøj C, Aviles-Rivero AI, *et al.* Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology. 2022 Jan;302(1):88-104. doi: 10.1148/radiol.2021210391.

(9) Sitio web de la FDA. Documento de autorización para MammoScreen 2.0. Publicado el 26 de noviembre de 2021. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K211541. [Consultada el 31/03/2023].

(10) Sitio web de la FDA. Documento de autorización para Genius AI Detection. Publicado el 6 de octubre de 2022. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K221449. [Consultada el 31/03/2023].

(11) Sitio web de la FDA. Documento de autorización para ProFound AI Software V3.0. Publicado el 12 de marzo de 2021. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K203822. [Consultada el 31/03/2023].

(12) Sitio web de la FDA. Documento de autorización para Transpara 1.7.0. Publicado el 30 de julio de 2021. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K210404. [Consultada el 31/03/2023].

(13) Sitio web de la FDA. Documento de autorización para Lunit INSIGHT MMG. Publicado el 7 de febrero de 2022. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K211678. [Consultada el 31/03/2023].

(14) Sitio web de la FDA. Documento de autorización para Saige-Dx. Publicado el 12 de mayo de 2022. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K220105. [Consultada el 31/03/2023].

(15) Sitio web de la FDA. Documento de autorización para cmTriage. Publicado el 8 de marzo de 2019. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K183285. [Consultada el 31/03/2023].

(16) Sitio web de la FDA. Documento de autorización para HealthMammo. Publicado el 16 de julio de 2020. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K200905. [Consultada el 31/03/2023].

(17) Sitio web de la FDA. Documento de autorización para CogNet Qm TRIAGE. Publicado el 29 de septiembre de 2022. Disponible en: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K220080. [Consultada el 31/03/2023].

(18) Shea B J, Reeves B C, 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 Sep 21;358:j4008. doi: 10.1136/bmj.j4008.

(19) Whiting PF, Rutjes AW, Westwood ME, *et al.* QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic

accuracy studies. Ann Intern Med 2011;155:529-36. doi:7326/0003-4819-155-8-201110180-00009

(20) Lopez de Argumedo M, Reviriego E, Gutierrez A, Bayon JC. Actualización del Sistema de Trabajo Compartido para Revisiones Sistemáticas de la Evidencia Científica y Lectura Critica (Plataforma FLC 3.0). Ministerio de Sanidad, Servicios Sociales e Igualdad. Servicio de Evaluación de Tecnologías Sanitarias del País Vasco; 2017. Informes de Evaluación de Tecnologías Sanitarias: OSTEBA.

(21) Bao C, Shen J, Zhang Y, Zhang Y, Wei W, Wang Z, Ding J, Han L. Evaluation of an artificial intelligence support system for breast cancer screening in Chinese people based on mammogram. Cancer Med. 2023 Feb;12(3):3718-3726. doi: 10.1002/cam4.5231.

(22) Dang LA, Chazard E, Poncelet E, Serb T, Rusu A, Pauwels X, Parsy C, Poclet T, Cauliez H, Engelaere C, Ramette G, Brienne C, Dujardin S, Laurent N. Impact of artificial intelligence in breast cancer screening with mammography. Breast Cancer. 2022 Nov;29(6):967-977. doi: 10.1007/s12282-022-01375-9.

(23) Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, Shin GW, Kim YJ, Choi HY. Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study. Korean J Radiol. 2022 May;23(5):505-516. doi: 10.3348/kjr.2021.0476.

(24) Sun Y, Qu Y, Wang D, Li Y, Ye L, Du J, Xu B, Li B, Li X, Zhang K, Shi Y, Sun R, Wang Y, Long R, Chen D, Li H, Wang L, Cao M. Deep learning model improves radiologists' performance in detection and classification of breast lesions. Chin J Cancer Res. 2021 Dec 31;33(6):682-693. doi: 10.21147/j.issn.1000-9604.2021.06.05.

(25) van Winkel SL, Rodríguez-Ruiz A, Appelman L, Gubern-Mérida A, Karssemeijer N, Teuwen J, Wanders AJT, Sechopoulos I, Mann RM. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol. 2021 Nov;31(11):8682-8691. doi: 10.1007/s00330-021-07992-w.

(26) Hsu W, Hippe DS, Nakhaei N, Wang PC, Zhu B, Siu N, Ahsen ME, Lotter W, Sorensen AG, Naeim A, Buist DSM, Schaffter T, Guinney J, Elmore JG, Lee CI. External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence. JAMA Netw Open. 2022 Nov 1;5(11):e2242343. doi: 10.1001/jamanetworkopen.2022.42343.

(27) Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health. 2022 Jul;4(7):e507-e519. doi: 10.1016/S2589-7500(22)00070-X.

(28) Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation. Radiology. 2022 Mar;302(3):535-542. doi: 10.1148/radiol.211590.

(29) Sharma N, Ng AY, James JJ, Khara G, Ambrozay E, Austin CC, *et al.* Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. medRxiv. 2021/252537. doi: https://doi.org/10.1101/2021.02.26.21252537.

(30) Larsen M, Aglen CF, Lee CI, Hoff SR, Lund-Hanssen H, Lång K, *et al.* Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. Radiology. 2022 Jun;303(3):502-11. doi: 10.1148/radiol.212381.

(31) Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, Lynge E, Vejborg I, Nielsen M, *et al.* An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology. 2022 Jul;304(1):41-9. doi: 10.1148/radiol.210948.

(32) Lotter W, Diab AR, Haslam B, *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med 2021;27:244-9. doi:1038/s41591-020-01174-9.

(33) McKinney SM, Sieniek M, Godbole V, *et al.* International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94. doi:1038/s41586-019-1799-6.

(34) Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst 2019;111:916-22. doi:1093/jnci/djy222.

(35) Salim M, Wåhlin E, Dembrower K, *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol 2020;6:1581-8. doi:1001/jamaoncol.2020.3321.

(36) Schaffter T, Buist DSM, Lee CI, *et al.* Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret

Screening Mammograms. JAMA Netw 2020;3:e200265. doi:1001/jamanetworkopen.2020.0265.

(37) Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. Radiol Artif Intell 2020;2:e190208. doi:1148/ryai.2020190208.

(38) Rodríguez-Ruiz A, Krupinski E, Mordang JJ, *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019;290:305-14. doi:1148/radiol.2018181371.

(39) Watanabe AT, Lim V, Vu HX, *et al.* Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging 2019;32:625-37. doi:1007/s10278-019-00192-5.

(40) Balta C, Rodriguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner SH. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? Proc SPIE 2020;11513:115130D. doi: 10.1117/12.2564179.

(41) Dembrower K, Wåhlin E, Liu Y, *et al.* Effect of artificial intelligencebased triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020;2:e468-74. doi:1016/S2589-7500(20)30185-0.

(42) Lang K, Dustler M, Dahlblom V, Akesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol 2021;31:1687-92. doi:1007/s00330-020-07165-1.

(43) Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. Radiology 2021;300:57-65. doi:1148/radiol.2021203555.

(44) Mital S, Nguyen HV. Cost-effectiveness of using artificial intelligence versus polygenic risk score to guide breast cancer screening. BMC Cancer. 2022 May 6;22(1):501. doi: 10.1186/s12885-022-09613-1.

(45) Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, *et al.* Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncol. 2016;2:1295-302. doi: 10.1001/jamaoncol.2016.1025.

(46) Vachon CM, Pankratz VS, Scott CG, Haeberle L, Ziv E, Jensen MR, *et al.* The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst. 2015;107:dju397.

(47) Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292:60-6. doi: 10.1148/radiol.2019182716.

(48) Naber SK, Kundu S, Kuntz KM, Dotson WD, Williams MS, Zauber AG, *et al.* Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk: current status and future potential. JNCI Cancer Spectr. 2020;4(1):pkz086. doi: 10.1093/jncics/pkz086.

(49) Kundu S, Kers JG, Janssens ACJ. Constructing hypothetical risk data from the area under the ROC curve: modelling distributions of polygenic risk. Plos One. 2016;11:e0152359. doi: 10.1371/journal.pone.0152359.

(50) Kerlikowske K, Hubbard RA, Miglioretti DL, Geller BM, Yankaskas BC, Lehman CD, *et al.* Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study. Ann Intern Med. 2011;155:493-502. doi: 10.7326/0003-4819-155-8-201110180-00005.

(51) European Society of Radiology. The cost of AI in radiology: is it really worth it? 2019. https://ai.myesr.org/healthcare/the-cost-of-ai-in-radiology-is-it-really-worth-it/.

(52) Stout NK, Lee SJ, Schechter CB, Kerlikowske K, Alagoz O, Berry D, *et al.* Benefits, harms, and costs for breast cancer screening after US implementation of digital mammography. J Natl Cancer Inst. 2014;106:dju092. doi: 10.1093/jnci/dju092.

(53) Shih Y-CT, Dong W, Xu Y, Shen Y. Assessing the cost-effectiveness of updated breast cancer screening guidelines for average-risk women. Value Health. 2019;22:185-93. doi: 10.1016/j.jval.2018.07.880.

(54) Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Ann Intern Med. 2011;155:10-20. doi: 10.7326/0003-4819-155-1-201107050-00003.

(55) Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018;286:800-9. doi:1148/radiol.2017171920.

(56) Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2020;14:49-58. doi:1093/ckj/sfaa188.

(57) Anderson AW, Marinovich ML, Houssami N, Lowry KP, Elmore JG, Buist DSM, Hofvind S, Lee CI. Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review. J Am Coll Radiol. 2022 Feb;19(2 Pt A):259-273. doi: 10.1016/j.jacr.2021.11.008.

(58) Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Br J Cancer. 2013 Jun 11;108(11):2205-40. doi: 10.1038/bjc.2013.177.

(59) Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer. 2017 Apr 13;3:12. doi: 10.1038/s41523-017-0014-x.

(60) Hovda T, Hoff SR, Larsen M, Romundstad L, Sahlberg KK, Hofvind S. True and missed interval cancer in organized mammographic screening: a retrospective review study of diagnostic and prior screening mammograms. Acad Radiol. 2022 Jan;29 Suppl 1:S180-S191. doi: 10.1016/j.acra.2021.03.022.

(61) Hofvind S, Skaane P, Vitak B, Wang H, Thoresen S, Eriksen L, Bjørndal H, Braaten A, Bjurstam N. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology. 2005 Nov;237(2):437-43. doi: 10.1148/radiol.2372041174.

(62) Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? 2021 Aug;31(8):5940-5947. doi: 10.1007/s00330-021-07686-3.

(63) Byng D, Strauch B, Gnas L, Leibig C, Stephan O, Bunk S, Hecht G. AI-based prevention of interval cancers in a national mammography screening program. Eur J Radiol. 2022 Jul;152:110321. doi: 10.1016/j.ejrad.2022.110321.

(64) Lang K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, Hofvind S, Andersson I, Rosso A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded,

screening accuracy study. Lancet Oncol. 2023 Aug;24(8):936-944. doi: 10.1016/S1470-2045(23)00298-X.

(65) Xunta de Galicia. Consellería de Sanidade. Expediente AB-CON3-23-015 para la contratación de un software basado en inteligencia artificial para apoyar a los radiólogos del PGDPCM en la lectura de las mamografías 3D y/o tomosíntesis, con un alcance de realización de un mínimo de 35.000 estudios de mamografía. [Consultado el 10 de enero de 2024]. Disponible en: https://extranet.sergas.es/cadmoweb/Cadmo_Web/DetalleContratacion.aspx?idPaxina=70003&idExpediente=77273

(66) Sociedad Española de Radiología Médica (SERAM). Cargas de trabajo en Radiología. SERAM/SEGECA julio 2020. [Consultado el 10 de enero de 2024]. Disponible en: https://seram.es/nuevo-documento-cargas-de-trabajo-radiologia/

(67) Decreto 351/2013, de 21 de mayo, de modificación del Decreto por el que se aprueba el Acuerdo regulador de las condiciones de trabajo del personal de Osakidetza-Servicio vasco de salud, para los años 2007, 2008 y 2009. BOPV (País Vasco), número 100, de 27 de mayo de 2013.

(68) Sociedad Española de Radiología Médica (SERAM). Catálogo de exploraciones de la SERAM edición 2016. [Consultado el 10 de enero de 2024]. Disponible en: https://seram.es/catalogo-seram/

(69) Sociedad Española de Radiología Médica (SERAM). ¿Cómo se mide la Actividad Radiológica en España y a nivel Internacional? [Consultado el 10 de enero de 2024]. Disponible en: https://piper.espacio-seram.com/index.php/seram/article/view/2799

(70) Vargas-Palacios A, Sharma N, Sagoo GS. Cost-effectiveness requirements for implementing artificial intelligence technology in the Women's UK Breast Cancer Screening service. Nat Commun. 2023 Sep 30;14(1):6110. doi: 10.1038/s41467-023-41754-0.

(71) Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast. 2020 Feb;49:25-32. doi: 10.1016/j.breast.2019.10.001.

(72) Lokaj B, Pugliese MT, Kinkel K, Lovis Ch, Schmid J. Barriers and facilitators of artificial intelligence conception and implementation for breast imaging diagnosis in clinical practice: a scoping review. Eur Radiol. 2023 Sept 02. doi: 10.1007/s00330-023-10181-6.

(73) Goisauf M, Cano Abadía M. Ethics of AI in Radiology: A Review of Ethical and Societal Implications. Front Big Data. 2022 Jul 14;5:850383. doi: 10.3389/fdata.2022.850383.

VI. Anexos

Anexo VI.1. Estrategias de búsqueda

VI.1.1. Búsqueda de revisiones sistemáticas y estudios primarios

Fecha de búsqueda: septiembre 2022.

Medline, vía Pubmed

Cáncer de mama/mamografía

#1 Search: "Breast Neoplasms"[Mesh]

#2 Search: (breast[Title/Abstract] OR mammary[Title/Abstract]) AND (neoplasm[Title/Abstract] OR neoplasms[Title/Abstract] OR tumor[Title/Abstract] OR tumors[Title/Abstract] OR tumour[Title/Abstract] OR tumours[Title/Abstract] OR cancer[Title/Abstract] OR cancers[Title/Abstract] OR carcinoma[Title/Abstract] OR carcinomas[Title/Abstract])

#3 Search: #1 OR #2 484,312

Inteligencia artificial

#4 Search: "Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh]

#5 Search: (artificial[Title/Abstract] OR computational[Title/Abstract] OR machine[Title/Abstract]) AND intelligence[Title/Abstract]

#6 Search: (machine[Title/Abstract] OR deep[Title/Abstract] OR transfer[Title/Abstract] OR hierarchical[Title/Abstract]) AND learning[Title/Abstract]

#7 Search: #4 OR #5 OR #6 227,630

Diagnóstico/cribado

#8 Search: "Mammography"[Mesh]

#9 Search: mammogra*[Title/Abstract] OR (breast[Title/Abstract] AND (tomosynthes*[Title/Abstract] OR ultrasonogra*[Title/Abstract]))

#10 Search: "Diagnosis"[Mesh] OR "Early Diagnosis"[Mesh] OR "Early Detection of Cancer"[Mesh]

#11 Search: diagnos*[Title/Abstract] OR detect*[Title/Abstract]

#12 Search: "Mass Screening"[Mesh]

#13 Search: screen*[Title/Abstract]

#14 Search: #8 OR #9 OR #10 OR #11 OR #12 OR #13 12,489,404

Medidas de resultado

#15 Search: "Sensitivity and Specificity"[Mesh] OR "Predictive Value of Tests"[Mesh] OR "ROC Curve"[Mesh] OR "Area Under Curve"[Mesh]

#16 Search: sensitivit*[Title/Abstract] OR specificit*[Title/Abstract] OR "predictive value"[Title/Abstract] OR (roc[Title/Abstract] AND (curve[Title/Abstract] OR analys*[Title/Abstract])) OR auc[Title/Abstract] OR auroc[Title/Abstract] OR auc-roc[Title/Abstract]

#17 Search: "Prognosis"[Mesh] OR prognos*[Title/Abstract]

#18 Search: "Reproducibility of Results"[Mesh] OR "Data Accuracy"[Mesh]

#19 Search: reproducibilit*[Title/Abstract] OR reliabilit*[Title/Abstract] OR validit*[Title/Abstract] OR accura*[Title/Abstract] OR performance[Title/Abstract] OR utilit*[Title/Abstract]

#20 Search: "Diagnostic Errors"[Mesh] OR "False Negative Reactions"[Mesh] OR "False Positive Reactions"[Mesh] OR "Observer Variation"[Mesh]

#21 Search: "false negative"[Title/Abstract] OR "false positive"[Title/Abstract] OR "true positive"[Title/Abstract] OR "true negative"[Title/Abstract] OR npv[Title/Abstract] OR ppv[Title/Abstract]

#22 Search: (observer[Title/Abstract] OR interobserver[Title/Abstract] OR inter-observer[Title/Abstract] OR intraobserver[Title/Abstract] OR intra-observer[Title/Abstract]) AND (variation[Title/Abstract] OR variations[Title/Abstract] OR bias[Title/Abstract])

#23 Search: #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 5,924,622

#24 Search: #3 AND #7 AND #14 AND #23 3,103

Revisiones/metaanalisis

#25 Search: #24 Filters: Meta-Analysis, Systematic Review 33

#26 Search: "Systematic Review"[Publication Type] OR "Systematic Reviews as Topic"[Mesh]

#27 Search: (systematic[Title/Abstract] OR evidence[Title/Abstract] OR literature[Title/Abstract]) AND (review*[Title/Abstract] OR overview*[Title/Abstract])

#28 Search: "Meta-Analysis"[Publication Type] OR "Meta-Analysis as Topic"[Mesh]

#29 Search: ("meta analy*"[Title/Abstract] OR metanaly*[Title/Abstract] OR metaanaly*[Title/Abstract] OR meta-analys*[Title/Abstract])

#30 Search: #26 OR #27 OR #28 OR #29 1,160,524

#31 Search: #24 AND #30 108

#32 Search: #25 OR #31 108

#33 Search: #32 Filters: English, Spanish **105**

Embase, vía OvidWeb

1 breast tumor/

2 (breast or mammary).ab,ti.

3 (neoplasm or neoplasms or tumor or tumors or tumour or tumours or cancer or cancers).ab,ti.

4 ((breast or mammary) adj2 (neoplasm or neoplasms or tumor or tumors or tumour or tumours or cancer or cancers)).ab,ti.

5 1 or 4 541807

6 artificial intelligence/

7 machine learning/

8 deep learning/

9	((artificial or computational or machine) adj2 intelligence).ab,ti.
10	((machine or deep or transfer or hierarchical) adj2 learning).ab,ti.
11	6 or 7 or 8 or 9 or 10 161599
12	mammography/
13	"mammogra*".ab,ti.
14	(breast adj2 (tomosynthes* or ultrasonogra*)).ab,ti.
15	early cancer diagnosis/ or diagnosis/ or early diagnosis/
16	(diagnos* or detect*).ab,ti.
17	mass screening/ or cancer screening/ or screening/
18	screen*.ab,ti.
19	12 or 13 or 14 or 15 or 16 or 17 or 18 8493444
20	5 and 11 and 19 2473
21	"sensitivity and specificity"/
22	(sensitivit* or specificit*).ab,ti.
23	predictive value/
24	"predictive value".ab,ti.
25	receiver operating characteristic/
26	"receiver operating characteristic".ab,ti.
27	(roc adj2 (curve or analys*)).ab,ti.
28	area under the curve/
29	(auc or auroc or auc-roc).ab,ti.
30	prognosis/ or cancer prognosis/
31	"prognos*".ab,ti.
32	reproducibility/
33	accuracy/
34	(reproducibilit* or reliabilit* or validit* or accura* or performance or utilit*).ab,ti.
35	diagnostic error/
36	((false or error) adj2 diagnos*).ab,ti.
37	false negative result/

38 false positive result/

39 ("false negative" or "false positive" or "true positive" or "true negative" or npv or ppv).ab,ti.

40 observer variation/

41 observer bias/

42 (observer or interobserver or inter-observer or intraobserver or intra-observer).ab,ti.

43 (variation or variations or bias).ab,ti.

44 ((observer or interobserver or inter-observer or intraobserver or intra-observer) adj2 (variation or variations or bias)).ab,ti.

45 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 44 6104445

46 20 and 45 2145

47 limit 46 to conference abstracts

48 46 not 47 1722

49 "systematic review"/ or "systematic review (topic)"/

50 meta analysis/ or "meta analysis (topic)"/

51 (systematic or evidence or literature).ab,ti.

52 (review* or overview*).ab,ti.

53 ((systematic or evidence or literature) adj2 (review* or overview*)).ab,ti.

54 ("meta analy*" or metanaly* or metaanaly* or meta-analys*).ab,ti.

55 49 or 50 or 53 or 54 881505

56 48 and 55 52

57 limit 56 to (english or spanish) **52**

Cochrane Library

#1 MeSH descriptor: [Breast Neoplasms] explode all trees

#2 (breast OR mammary):ti,ab,kw AND (neoplasm OR neoplasms OR tumor OR tumors OR tumour OR tumours OR cancer OR cancers OR carcinoma OR carcinomas):ti,ab,kw

#3	#1 OR #2	42973
#4	MeSH descriptor: [Artificial Intelligence] explode all trees	
#5	MeSH descriptor: [Machine Learning] explode all trees	
#6	MeSH descriptor: [Deep Learning] explode all trees	
#7	(artificial OR computational OR machine):ti,ab,kw AND (intelligence):ti,ab,kw	
#8	(machine OR deep OR transfer OR hierarchical):ti,ab,kw AND (learning):ti,ab,kw	
#9	#4 OR #5 OR #6 OR #7 OR #8	6204
#10	MeSH descriptor: [Mammography] explode all trees	
#11	(mammogra*):ti,ab,kw	
#12	(breast):ti,ab,kw AND (tomosynthes* OR ultrasonogra*):ti,ab,kw	
#13	MeSH descriptor: [Diagnosis] explode all trees	
#14	MeSH descriptor: [Early Diagnosis] explode all trees	
#15	MeSH descriptor: [Early Detection of Cancer] explode all trees	
#16	(diagnos* OR detect*):ti,ab,kw	
#17	MeSH descriptor: [Mass Screening] explode all trees	
#18	(screen*):ti,ab,kw	
#19	#10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18	632590
#20	#3 AND #9 AND #19	**78**
	RS	**0**

Web of Science

1	(TI=(breast OR mammary)) OR AB=(breast OR mammary)	
2	(TI=(neoplasm OR neoplasms OR tumor OR tumors OR tumour OR tumours OR cancer OR cancers OR carcinoma OR carcinomas)) OR AB=(neoplasm OR neoplasms OR tumor OR tumors OR tumour OR tumours OR cancer OR cancers OR carcinoma OR carcinomas)	
3	#2 AND #1	628,801

4 (TI=(artificial OR computational OR machine)) OR AB=(artificial OR computational OR machine)

5 (TI=(intelligence)) OR AB=(intelligence)

6 #5 AND #4

7 (TI=(machine OR deep OR transfer OR hierarchical)) OR AB=(machine OR deep OR transfer OR hierarchical)

8 (TI=(learning)) OR AB=(learning)

9 #8 AND #7

10 #9 OR #6 565,487

11 (TI=(mammogra*)) OR AB=(mammogra*)

12 (TI=(breast)) OR AB=(breast)

13 (TI=(tomosynthes* OR ultrasonogra*)) OR AB=(tomosynthes* OR ultrasonogra*)

14 #12 AND #13

15 (TI=(diagnos* OR detect* OR screen*)) OR AB=(diagnos* OR detect* OR screen*)

16 #14 OR #11 OR #15 11,143,280

17 #16 AND #10 AND #3 3,345

18 (TI=(sensitivit* or specificit*)) OR AB=(sensitivit* or specificit*)

19 (TI=("predictive value")) OR AB=("predictive value")

20 (TI=("receiver operating characteristic")) OR AB=("receiver operating characteristic")

21 (TI=(roc AND (curve or analys*))) OR AB=(roc AND (curve or analys*))

22 (TI=(auc or auroc or auc-roc)) OR AB=(auc or auroc or auc-roc)

23 (TI=(prognos*or reproducibilit* or reliabilit* or validit* or accura* or performance or utilit*)) OR AB=(prognos*or reproducibilit* or reliabilit* or validit* or accura* or performance or utilit*)

24 (TI=((false or error) and diagnos*)) OR AB=((false or error) and diagnos*)

25 (TI=("false negative" or "false positive" or "true positive" or "true negative" or npv or ppv)) OR AB=("false negative" or "false positive" or "true positive" or "true negative" or npv or ppv)

26 (TI=(observer or interobserver or inter-observer or intraobserver or intra-observer)) OR AB=(observer or interobserver or inter-observer or intraobserver or intra-observer)

27 (TI=(variation or variations or bias)) OR AB=(variation or variations or bias)

28 #27 AND #28

29 #28 OR #25 OR #24 OR #23 OR #22 OR #21 OR #20 OR #19 OR #18 12,418,465

30 #17 AND #29 2,719

31 (TI=(systematic or evidence or literature)) OR AB=(systematic or evidence or literature)

32 (TI=(review* or overview*)) OR AB=(review* or overview*)

33 #32 AND #31

34 (TI=("meta analy*" or metanaly* or metaanaly* or meta-analys*)) OR AB=("meta analy*" or metanaly* or metaanaly* or meta-analys*)

35 #34 OR #33 1,650,027

36 #35 AND #30 76
Refined By:Document Types: Article or Review Article or Unspecified or Other or Early Access. Languages: English **73**

Scopus

1 (TITLE-ABS-KEY (breast OR mammary) AND TITLE-ABS-KEY (neoplasm OR neoplasms OR tumor OR tumors OR tumour OR tumours OR cancer OR cancers OR carcinoma OR carcinomas))
672,301 results

2 (TITLE-ABS-KEY (artificial OR computational OR machine) AND TITLE-ABS-KEY (intelligence))

3 (TITLE-ABS-KEY (machine OR deep OR transfer OR hierarchical) AND TITLE-ABS-KEY (learning))

4 3 OR 2
1,082,936 results

5 TITLE-ABS-KEY (mammogra*)

6 (TITLE-ABS-KEY (breast) AND TITLE-ABS-KEY (tomosynthes* OR ultrasonogra*))

7 TITLE-ABS-KEY (diagnos* OR detect* OR screen*)

8 7 OR 6 OR 5
11,879,073 results

9 8 AND 4 AND 1
7,077 results

10 (TITLE-ABS-KEY (sensitivit* OR specificit*) OR TITLE-ABS-KEY (“predictive value”) OR TITLE-ABS-KEY (“receiver operating characteristic”) OR TITLE-ABS-KEY (roc AND (curve OR analys*)) OR TITLE-ABS-KEY (auc OR auroc OR auc-roc) OR TITLE-ABS-KEY (prognos*or AND reproducibilit* OR reliabilit* OR validit* OR accura* OR performance OR utilit*) OR TITLE-ABS-KEY ((false OR error) AND diagnos*) OR TITLE-ABS-KEY (“false negative” OR “false positive” OR “true positive” OR “true negative” OR npv OR ppv))

11 (TITLE-ABS-KEY (observer OR interobserver OR inter-observer OR intraobserver OR intra-observer) AND TITLE-ABS-KEY (variation OR variations OR bias))

12 11 OR 10
3,918,050 results

13 12 AND 9
3,088 results

14 (TITLE-ABS-KEY (systematic OR evidence OR literature) AND TITLE-ABS-KEY (review* OR overview*))

15 TITLE-ABS-KEY (“meta analy*” OR metanaly* OR metaanaly* OR meta-analys*)

16 15 OR 14
2,002,190 results

17 16 AND 13
113 results

18 17 AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (SRCTYPE , “j”)) **106 results**

VI.1.2. Búsqueda de estudios de costes/económicos

Fecha de búsqueda: diciembre 2022.

NHS EED, CEA Registry, IHE

((breast OR mammary) AND (cancer* OR neoplasm* OR tumor* OR tumours*)) AND ((artificial AND intelligence) OR (machine AND learning)) AND (mammogram* OR screen*) 0

Medline, vía Pubmed

Cáncer de mama/mamografía

#1 Search: "Breast Neoplasms"[Mesh]

#2 Search: (breast[Title/Abstract] OR mammary[Title/Abstract]) AND (neoplasm[Title/Abstract] OR neoplasms[Title/Abstract] OR tumor[Title/Abstract] OR tumors[Title/Abstract] OR tumour[Title/Abstract] OR tumours[Title/Abstract] OR cancer[Title/Abstract] OR cancers[Title/Abstract] OR carcinoma[Title/Abstract] OR carcinomas[Title/Abstract])

#3 Search: #1 OR #2 491,504

Inteligencia artificial

#4 Search: "Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh]

#5 Search: (artificial[Title/Abstract] OR computational[Title/Abstract] OR machine[Title/Abstract]) AND intelligence[Title/Abstract]

#6 Search: (machine[Title/Abstract] OR deep[Title/Abstract] OR transfer[Title/Abstract] OR hierarchical[Title/Abstract]) AND learning[Title/Abstract]

#7 Search: #4 OR #5 OR #6 242,198

Diagnóstico/cribado

#8 Search: "Mammography"[Mesh]

#9 Search: mammogra*[Title/Abstract] OR (breast[Title/Abstract] AND (tomosynthes*[Title/Abstract] OR ultrasonogra*[Title/Abstract]))

#10 Search: "Diagnosis"[Mesh] OR "Early Diagnosis"[Mesh] OR "Early Detection of Cancer"[Mesh]

#11 Search: diagnos*[Title/Abstract] OR detect*[Title/Abstract]

#12 Search: "Mass Screening"[Mesh]

#13 Search: screen*[Title/Abstract]

#14 Search: #8 OR #9 OR #10 OR #11 OR #12 OR #13 12,636,346

#15 Search: #3 AND #7 AND #14 3,894

Costes

#16 Search: "Economics"[Mesh]

#17 Search: "Models, Economic"[Mesh]

#18 Search: "Costs and Cost Analysis"[Mesh] OR "Cost Allocation"[Mesh] OR "Cost-Benefit Analysis"[Mesh] OR "Cost Control"[Mesh] OR "Health Care Costs"[Mesh] OR "Health Expenditures"[Mesh]

#19 Search: "Decision Trees"[Mesh] OR "Monte Carlo Method"[Mesh] OR "Markov Chains"[Mesh]

#20 Search: cost*[Title/Abstract] OR economic*[Title/Abstract] OR pric*[Title/Abstract] OR expenditure*[Title/Abstract] OR fee[Title/Abstract] OR fees[Title/Abstract] OR modelization[Title/Abstract]

#21 Search: decision*[Title/Abstract] AND (tree*[Title/Abstract] OR analys*[Title/Abstract])

#22 Search: decision-tree*[Title/Abstract]

#23 Search: financial[Title/Abstract] AND impact*[Title/Abstract]

#24 Search: "monte carlo"[Title/Abstract] OR markov[Title/Abstract]

#25 Search: #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 1,769,410

#26 #15 AND #25 599

#27 #26 Filters: from 2012-2023 497

#28 Search: #27 Filters: English, Spanish **495**

Embase, via OvidWeb

1 breast tumor/

2 (breast or mammary).ab,ti.

3 (neoplasm or neoplasms or tumor or tumors or tumour or tumours or cancer or cancers).ab,ti.

4 ((breast or mammary) adj2 (neoplasm or neoplasms or tumor or tumors or tumour or tumours or cancer or cancers)).ab,ti.

5 1 or 4 549761

6 artificial intelligence/

7 machine learning/

8 deep learning/

9 ((artificial or computational or machine) adj2 intelligence).ab,ti.

10 ((machine or deep or transfer or hierarchical) adj2 learning).ab,ti.

11 6 or 7 or 8 or 9 or 10 177186

12 mammography/

13 "mammogra*".ab,ti.

14 (breast adj2 (tomosynthes* or ultrasonogra*)).ab,ti.

15 early cancer diagnosis/ or diagnosis/ or early diagnosis/

16 (diagnos* or detect*).ab,ti.

17 mass screening/ or cancer screening/ or screening/

18 screen*.ab,ti.

19 12 or 13 or 14 or 15 or 16 or 17 or 18 8635182

20 5 and 11 and 19 2712

47 economics/ or health economics/

48 "cost utility analysis"/ or "cost benefit analysis"/ or "health care cost"/ or "cost"/ or "cost effectiveness analysis"/ or "cost control"/ or "program cost effectiveness"/

49 (economic* or cost* or pric* or expenditur* or expens* or financ* or fee or fees or modelization).ab,ti.

50 economic model/

51 "decision tree"/

52 Monte Carlo method/

53 Markov chain/

54 (decision* and (tree* or analys*)).ab,ti.

55 "decision-tree*".ab,ti.

56 (financial and impact*).ab,ti.

57 (“monte carlo” or markov).ab,ti.

58 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 2275948

59 46 and 58 572

60 limit 59 to conference abstracts

61 59 not 60 435

62 limit 61 to yr=”2012 -Current” 405

63 limit 62 to (english or spanish) **403**

Anexo VI.2. QUADAS-2

VI.2.1. Adaptación del cuestionario QUADAS-2

Item	Respuesta
SELECCIÓN DE PARTICIPANTES - A. RIESGO DE SESGO	
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	**Sí** – ECAs y estudios de cohortes (prospectivos o retrospectivos) con muestra no enriquecida (consecutiva o aleatorizada). **Poco claro** – Si no se indica. **No** – Otros estudios.
¿Evitó el estudio exclusiones inapropiadas?	**Sí** – Si se evitaron exclusiones inapropiadas. **Poco claro** – Si no se notifica claramente. **No** – Exclusión de más del 10 % de las muestras por cualquier razón (p. ej., estudios retrospectivos con datos perdidos). **No** – Exclusión sistemática de tipos de mujeres / imágenes (p. ej., de mama densa). **No** – Exclusión basada en resultados (p. ej., exclusión de tipos de cáncer, exclusión de cánceres de intervalo, exclusión / inclusión basada en una decisión de revaloración).
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	En el caso de los estudios de conjuntos de pruebas, esto se traduce en: ¿el conjunto de pruebas se ha descrito claramente como un conjunto de validación externo (geográficamente)? **Sí** – Validación geográfica externa (el conjunto de pruebas era una muestra de un centro diferente; puede estar en otro país o en el mismo país). **Poco claro** – No se dan detalles sobre el conjunto de entrenamiento y el conjunto de ajuste. **No** – Cualquier validación interna (p. ej., muestras divididas, validación cruzada) o validación temporal. Para estudios prospectivos aplicados en un contexto clínico: **Sí** – Si el estudio se localiza en diferente(s) centro(s) a aquellos que proporcionaron las mamografías utilizadas en el entrenamiento y ajuste del algoritmo de IA. **Poco claro** – Si no se indica. **No** – Si hay algún solapamiento.
SELECCIÓN DE PARTICIPANTES - B. PREOCUPACIONES RESPECTO A LA APLICABILIDAD	
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	**Alto** – Si «sí» para alguna de las siguientes afirmaciones. **Poco claro** – Si no se proporcionan detalles. **Bajo** – Si «no» para todas de las siguientes afirmaciones. • No es una muestra consecutiva o aleatoria de las mujeres que acuden al cribado. • Muestra enriquecida / la prevalencia del cáncer no coincide con el contexto del cribado (> 3 %). • Mamografías no procedentes de DM o BDT de campo completo. • Mamografías que no proceden del cribado (p. ej., diagnósticas o sintomáticas) o sólo se incluye un subconjunto, como los casos revalorados o los FN (el cáncer puede ser más fácil o difícil de detectar). Mamografías de mujeres no representativas de la población española (raza, edad).

Item	Respuesta
PRUEBA ÍNDICE - A. RIESGO DE SESGO	
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Para las pruebas de índice en las que interviene un ser humano, ya sea un comparador de lectura humano (IA como ayuda a la lectura) o incluido de otro modo en la vía de pruebas de IA (arbitraje): **Sí** – Requiere una declaración clara de cegamiento, o relaciones temporales claras donde la lectura humana se haya producido antes que el patrón de referencia. **No** – En caso contrario. Para la prueba índice en donde la IA es utilizada sin ningún elemento humano: **Sí** – El sistema de IA no ha sido entrenado previamente con estas mamografías ni ha aprendido de ellas ni de otras mamografías de las mismas mujeres. **No** – Si se repite el uso de los mismos casos (a menos que se explicite que el algoritmo de IA estaba preestablecido y no cambió al repetirse el uso, y el estudio no seleccionó uno de los varios sistemas de IA basándose en el uso con los mismos casos). **Poco claro** – Si no se explicita que ha sido repetido en el mismo estudio o en estudios anteriores.
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	**No** – Si los lectores humanos no estuvieron cegados a la IA (a menos que la IA sea parte específica de la misma prueba índice). **No** – Si los sistemas de IA entrenan y calibran utilizando decisiones de los lectores humanos en los mismos casos. **Sí** – En caso contrario.
Si se utilizó un umbral ¿se especificó previamente?	**Sí** – Si utiliza un sistema de IA comercialmente disponible que da un resultado de si / no, o un umbral claramente preespecificado en los métodos. **Sí** – Para los sistemas que dan una puntuación de riesgo y el estudio establece explícitamente el umbral preespecificado. **No** – Utiliza la sensibilidad / especificidadS del lector como referencia utilizando el mismo conjunto de datos. **No** – Establece el umbral con el conjunto de validación sin evidencia temporal (p. ej., protocolo publicado) de que el umbral estaba realmente preespecificado. **No** – Lectores humanos o combinaciones de humano y IA.
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	**Sí** – Si los lectores tomaron las decisiones en un contexto clínico, y esas decisiones fueron utilizadas para decidir si revalorar o no a las mujeres (ya sea prospectivamente como parte del ensayo o estudio de precisión de la prueba o retrospectivamente utilizando la decisión original). **No** – Si los lectores examinaron un conjunto de pruebas (de cualquier prevalencia) fuera de la práctica clínica, o cualquier otro contexto susceptible de provocar el efecto laboratorio.
PRUEBA ÍNDICE - B. PREOCUPACIONES RESPECTO A LA APLICABILIDAD	
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	**Alto** – Si «sí» para alguna de las siguientes afirmaciones. **Poco claro** – Si no se proporcionan detalles. **Bajo** – Si «no» para alguna de las siguientes afirmaciones. • El sistema de IA no está todavía comercialmente disponible, p. ej., sistemas *in-house*; • El estudio no utilizó un umbral previamente especificado para el sistema de IA; • No es una vía de prueba completa aplicable a la práctica clínica (p. ej., precisión de la IA para una sola lectura, pero no está integrada en las decisiones del centro de cribado (el arbitraje)); • El comparador humano no es una vía de prueba completa aplicable a la práctica clínica (doble lectura humana con arbitraje en el umbral clínico); El sistema de IA / lector no tenía acceso a mamografías anteriores / no había cuatro vistas disponibles.

Item	Respuesta
PATRÓN DE REFERENCIA - A. RIESGO DE SESGO	
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	**Sí** – Si el patrón de referencia es el resultado histopatológico de la biopsia (cáncer presente o ausente) con al menos dos años de seguimiento para los cánceres de intervalo. **No** – Si el patrón de referencia es el resultado histopatológico de la biopsia (cáncer presente o ausente) sin seguimiento.
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	**Sí** – Estudios retrospectivos en donde los lectores leen las mamografías prospectivamente (conjunto de pruebas enriquecidas). **No** – Para estudios retrospectivos (si incluimos el comparador de lectores humanos como prueba índice). **No** – Para estudios prospectivos si los investigadores no cegaron a los clínicos que realizaban las pruebas de seguimiento con respecto a qué prueba índice examinaban las mamografías, p. ej., poniendo marcas de localización en el mismo formato para los lectores de IA y humanos.
PATRÓN DE REFERENCIA - B. PREOCUPACIONES RESPECTO A LA APLICABILIDAD	
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	**Alto** – Si «sí» para alguna de las siguientes afirmaciones. **Poco claro** – Si no se proporcionan detalles. **Bajo** – Si «no» para todas las siguientes afirmaciones. • Duración de las rondas del cribado < 2 años de seguimiento / definición de los cánceres de intervalo; Clasificación no mediante biopsia / seguimiento.
FLUJO Y TIEMPO - A. RIESGO DE SESGO	
¿Recibieron todos los pacientes un patrón de referencia?	**No** – Si hubo una pérdida significativa (> 10 %) durante el seguimiento para el patrón de referencia de los cánceres de intervalo o los resultados de cribado posteriores. **No** – Si cualquier mujer que debería haber recibido una biopsia o pruebas de seguimiento después de resultados positivos de la prueba índice no la recibió o no se dispuso de los resultados. **Sí** – En caso contrario.
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	**Sí** – Para los ECA de prueba-tratamiento aleatorizados a diferentes estrategias de prueba y sus decisiones de revaloración asociadas. **Sí** – Si las todas las mujeres que dan positivo en alguna de las pruebas índice incluidas (vías de IA o vías humanas de comparación) reciben pruebas de seguimiento / biopsia en un estudio prospectivo. **No** – Si las mujeres fueron revaloradas para más pruebas con base en una de las pruebas índice, y no de otra(s), esto causará sesgo porque el cáncer, cuando se presenta, es más probable que se detecte en personas que reciben pruebas de seguimiento después de la revaloración del cribado. **No** – En estudios retrospectivos, la decisión de volver a revalorar para realizar pruebas de seguimiento / biopsia se tomó en función de la decisión de los lectores humanos. No sabemos si los positivos de la IA y los negativos de los lectores humanos son FP o VP, ni qué tipo de VP. El seguimiento hasta el desarrollo de cánceres de intervalo detectará algunos de estos cánceres, pero no todos, por lo que reduce, pero no elimina, este sesgo. **No** – Para estudios prospectivos en donde la decisión de revalorar está informada por una prueba índice pero no por todas, o está más influenciada por una de las pruebas que por otras. **Poco claro** – Estudios de lectores retrospectivos (estudios de conjuntos de pruebas enriquecidos) en los que los lectores leen prospectivamente datos retrospectivos, el patrón de referencia no se basa en ninguna prueba índice, sino que el patrón de referencia se basa en la decisión original del lector humano. Los revisores no tienen claro el riesgo de sesgo.
¿Se incluyó a todos los pacientes en el análisis?	**Sí** – Si no hubo exclusiones después del momento de la selección de la cohorte, por ejemplo, resultados intermedios o indeterminados. **No** – En caso contrario.

VI.2.2. Resultados cuestionario QUADAS-2

Autor: Bao *et al*., 2022	
Características del estudio	
Muestra de pacientes	Estudio retrospectivo. En total se recogen 640 mamografías (edad media: 54 años; mujeres: 100 %; cáncer: 62,05 %), 321 evaluadas inicialmente por la IA. El 75 % de las mamografías evaluadas con ayuda de la IA y sin ayuda de la IA son leídas por dos especialistas en radiología, por lo que el número de mamografías evaluadas es de 564 en el grupo sin ayuda de la IA y 562 en el grupo con ayuda.
Características de los pacientes y ámbito	
	Criterios de inclusión Formato de imagen: imagen digital y comunicaciones en medicina (DICOM); en la imagen podría encontrarse una lesión; con sistema inicial de BI-RADS.
	Criterios de exclusión Mamografías de baja calidad; mamografías compuestas con menos de dos proyecciones (p. ej., vista CC izquierda, vista OML izquierda, vista CC derecha, vista OML derecha); lesión no identificable por la mamografía y se necesitó otro medio (p. ej., ultrasonido, imagen de resonancia magnética); sin resultado de biopsia.
	Ámbito clínico Dos hospitales 3A y un hospital terciario en China.

Prueba índice	El sistema de apoyo de IA es un software auxiliar de diagnóstico de mamografías (versión 3.2.3, Yizhun AI). El sistema se basa principalmente en el modelo de red neuronal profunda, indica a los especialistas en radiología el área, la categoría y los BI-RADS de las lesiones sospechosas para aumentar efectivamente la sensibilidad y evitar diagnósticos erróneos en el cribado de cáncer de mama. El sistema está entrenado y validado a partir de una base de datos de más de 16.000 mamografías con diferente tipo de lesiones (un tercio de ellas fueron calcificaciones y masas) y mamografías normales. Estas mamografías procedían de dispositivos de proveedores diferentes (GE, Philips, Siemens, Hologic, etc.). El sistema está probado en un conjunto de datos interno independiente de varios proveedores, no utilizado para el entrenamiento o la validación del algoritmo. Las mamografías usadas en este estudio no fueron nunca utilizadas para entrenar, validar o probar los algoritmos. El especialista en radiología utilizó una decisión interactiva al utilizar este sistema de apoyo en práctica. El sistema da una lista de lesiones correspondiente al caso analizado. El especialista en radiología clica en la lista de lesiones para visualizar el contorno de la lesión analizada e información del diagnóstico. El especialista en radiología puede releer la mamografía con base en la evaluación del sistema. Después de la relectura, el especialista en radiología revisa las características de las lesiones y BI-RADS si es necesario.
Patrón de referencia	Resultado de la biopsia.
Flujo y tiempo	Cuatro mamografías, dos en el grupo sin y dos en el grupo con ayuda de IA se excluyeron del examen al ser clasificadas por los especialistas en radiología como BI-RADS 0 (las lesiones no se pudieron identificar en las mamografías y se necesitaron otro tipo de exámenes). Lecturas totales: 562 sin y 560 con ayuda de IA.

Comparador	Mamografías evaluadas por especialistas en radiología con ayuda de la IA frente a sin ayuda de la IA. Las mamografías se dividen entre los dos grupos mediante una estratificación aleatorizada basada en los BI-RADS iniciales. Los especialistas en radiología juzgan si aceptan los consejos de la IA en relación con cada detalle (densidad de la mama, tipo de lesión). Los especialistas en radiología estuvieron cegados a cualquier información (p. ej., edad, sexo, resultado de la biopsia) relacionada con la persona cuya mamografía fue analizada. De cada grupo los especialistas en radiología registran la puntuación BI-RADS de 0 a 5 (el BI-RADS 4 incluye la puntuación 4A, 4B y 4C), el tipo de lesión y densidad de la mama.
Notas	643 mamografías examinadas, 322 sin y 321 con ayuda de IA. El 75 % de estas en cada grupo fueron leídas por dos especialistas en radiología. Lecturas totales: 564 sin y 562 con ayuda de IA

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto

DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	No		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	No		
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	No		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	Sí		
		Alto	Alto

DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	Poco claro		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Poco claro	
Autor: Dang *et al*., 2022			
Características del estudio			
Muestra de pacientes	Casos múltiples con lector múltiple y diseño cruzado. Datos recogidos retrospectivamente. Muestra enriquecida con casos de cáncer, los lectores desconocen en qué proporción. 397 exámenes de cribado incluidos inicialmente, 329 cumplen con los criterios de inclusión. Después de excluir aleatoriamente 15 casos benignos, se incluyeron 314 exámenes. Los exámenes se adquirieron con el sistema Hologíc Selenia 3D Dimension.		
Características de los pacientes y ámbito			
	Criterios de inclusión Mujeres entre 50 y 74 años, asintomáticas, sin historia personal o familiar de cáncer de mama o de cirugía de mama y sin factores genéticos de riesgo. Mamografía de 4 vistas estándar.		
	Criterios de exclusión		
	Ámbito clínico Un hospital en Francia.		

Prueba índice	Sistema de IA Mammoscreen v.1.2 (Therapixel, French) diseñado para detectar áreas sospechosas de contener cáncer de mama, para evaluar su grado de sospecha en mamografías digitales 2D. El sistema toma como input la vista CC y OML de cada mama y proporciona como resultado la posición de las lesiones sospechosas con una puntuación de sospecha para cada una de ellas clasificadas entre 1 y 10. El sistema fue aprobado por la FDA e 2020 y recibió el marcado CE en enero 2021. El sistema de IA se valida externamente mediante un estudio que es una investigación de casos múltiples con lector múltiple y diseño cruzado. Dos sesiones de lectura, retrasadas por un periodo de lavado de cuatro semanas. Durante cada periodo la mitad de los grupos de datos se leen con soporte de la IA y la otra mitad sin. 12 especialistas en radiología participan en el estudio. El orden de lectura se aleatorizó entre los especialistas en radiología los cuales no recibieron información adicional sobre las imágenes y los pacientes. De las 314 imágenes, para 85 exámenes se dispone de mamografías adquiridas previamente. Para cada caso los lectores: marcan la lesión más sospechosa para cada mama en la vista CC y OML, cuando es posible; asignan un BI-RADS entre 1 y 5 para cada lesión; y asignan un nivel de sospecha o «BI-RADS 100 continuo» definido de la siguiente manera: una escala que va de 1 a 100 (1-20 para BI-RADS 1, 21-40 para BI-RADS 2, 41-60 para BI-RADS 3, 61-80 para BI-RADS 4 y 81-00 para BIRADS 5). Para los exámenes leídos con la ayuda de la IA, los especialistas en radiología pueden comprobar la puntuación de sospecha asignada por la IA antes de reportar su evaluación.
Patrón de referencia	Definición estándar: casos de cáncer confirmados mediante biopsia positiva y cáncer negativo verificado mediante seguimiento negativo. Definición basada en experto: la clasificación BI-RADS asignada por un especialista en radiología experta durante la fase de inclusión.
Flujo y tiempo	Se excluyeron aleatoriamente 15 exámenes benignos de los 329 cumplen con los criterios de inclusión.

Comparador	Especialista en radiología asistido con la IA frente a especialista en radiología no asistido con la IA. Se realizan dos sesiones de lectura retrasadas por un periodo de descanso de cuatro semanas. En cada sesión la mitad de los grupos de datos se leen por especialistas en radiología con y sin ayuda de IA. El orden de lectura se aleatoriza entre los participantes. Los lectores no disponen de información adicional como imágenes adicionales o información sobre los pacientes. De los 314 exámenes incluidos para 85 se disponen de mamografías adquiridas previamente.
Notas	Se analizan 314 exámenes por cada mama, total 628, de los que el 20 % (128/628) son casos positivos de cáncer.

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	No		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Poco claro		
Si se utilizó un umbral ¿se especificó previamente?	No		

Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	No		
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	No		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Si		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	Poco claro		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Poco claro	

Autor: Lee *et al*. 2022	
Características del estudio	
Muestra de pacientes	Estudio retrospectivo con un total de 200 casos (100 casos malignos probados por biopsia, 40 benignos y 60 casos negativos). Los casos malignos se recogen consecutivamente, y 20 casos benignos y 60 negativos aleatoriamente.
Características de los pacientes y ámbito	
	Criterios de inclusión Mujeres asiáticas adultas a las que se realiza FFDM de cuatro vistas para cribado de cáncer de mama (adquiridas con Senographe 2000D, GE Healthcare).
	Criterios de exclusión Imagen de mala calidad. Implante mamario y marcapasos. Historia de cáncer de mama. Historia de cirugía de mama.
	Ámbito clínico Los datos se recogen en el Hospital Universitario Soonchunhyand de Bucheon, Corea.
Prueba índice	Sistema de IA Lunit INSIGHT MMG, versión 1.1.1.0 (Lunit), comercialmente disponible, basado en aprendizaje profundo y entrenado con 170.230 mamografías para la detección de cáncer de mama (30.000 casos de cáncer). El algoritmo de IA proporciona una puntuación de anormalidad basada en cuatro vistas entre 1 y 100, indicando la posibilidad de cáncer de mama y un mapa de calor en la ubicación de la región anormal, indicando una puntuación de anormalidad de 10 o más.
Patrón de referencia	Casos malignos probados: cáncer de mama patológicamente confirmado en los seis meses siguientes a la mamografía (no se incluye cáncer de mama bilateral). Casos negativos: mamografías con BI-RADS categoría 1 y confirmado como negativo durante más de dos años de seguimiento. Casos benignos: confirmados por biopsia o imagen de seguimiento durante más de dos años.

Flujo y tiempo	
Comparador	Cinco lectores (tres especialistas en radiología especialistas en mama y dos especialistas en radiología generales), leen las mamografías con la IA y después de dos meses de descanso, las leen sin la IA. Cinco lectores restantes leen las mamografías con y sin asistencia de la IA. No hay modificación en el ambiente de lectura entre sesiones de lectura con y sin IA. Los especialistas en radiología valoran la probabilidad de malignidad de las mamografías en cada caso en una escala de 7 puntos, tanto para la lectura con IA o sin IA: 1 = definitivamente normal, 2 = benigna, 3 = probablemente benigna, 4 = baja sospecha de malignidad, 5 = sospecha moderada de malignidad, 6 = sospecha alta de malignidad y 7 = altamente sugestivo de malignidad. Para que una lesión sea sospechosa de cáncer de mama la puntuación tiene que ser como mínimo de 3. Los lectores consideran tanto los resultados de la IA como los hallazgos mamográficos originales y hacen sus juicios finales utilizando la escala de 7 puntos.
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto

DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	No		
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Poco claro		

¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	Poco claro		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Poco claro	

Autor: Sun *et al*. 2021	
Características del estudio	
Muestra de pacientes	Estudio multicéntrico que incluye un diseño retrospectivo y prospectivo. Este se compone de tres partes: una 1.ª (retrospectiva, 4.119 mujeres) en la que se construye, entrena y valida internamente el sistema de IA; una 2.ª (200 mujeres inscritas retrospectivamente y seleccionadas al azar, 70 con cáncer) en la que se valida externamente; y una 3.ª (prospectiva, 5.746 mujeres inscritas consecutivamente, 832 con cáncer) en la que se verifica el efecto del sistema de IA en la práctica clínica.
Características de los pacientes y ámbito	
	Criterios de inclusión 1.ª parte: mujeres a las que se realiza mamografía en cada centro, con datos clínicos completos, datos mamográficos y con diagnóstico patológico o más de dos años de seguimiento después del primer examen. 2.ª parte: mujeres a las que se realiza mamografía en cada centro, con diagnóstico patológico o más de dos años de seguimiento después del primer examen. 3.ª parte: mujeres a las que se realiza mamografía en cada centro. Mamografías almacenadas en formato de imagen digital (dos vistas estándar, CC u OML) y DICOM.

	Criterios de exclusión 1.ª parte: imágenes incalificables, inconsistencia en la localización de la lesión entre las mamografías y los resultados patológicos. 2.ª parte: mujeres con síntomas extremadamente obvios de cáncer en la mamografía y con historia de cáncer de mama o diagnóstico claro en el momento de la visita. 3.ª parte: mujeres sin resultados patológicos, perdidas en el seguimiento de dos años o con mamografías de calidad insuficiente.
	Ámbito clínico Seis centros de Pekín, China.
Prueba índice	Sistema de IA *in-house* que contiene tres modelos neuronales profundos: módulo de detección, módulo de concordancia y módulo de evaluación del grado de malignidad de las lesiones. Para la detección de lesiones sospechosas en todas las imágenes de un paciente utiliza Faster R-CNN, Res-Net-50 como red básica y adopta las características de una red piramidal. Para indicar si un par de candidatos detectados provienen de diferentes vistas de la misma lesión, se utiliza un modelo neuronal. Para estimar el grado de malignidad de las lesiones usa una red neuronal convolucional ResNet. Las lesiones se representan mediante BI-RADS que se utilizan para entrenar el modelo (BI-RADS 1 y 2, BI-RADS 3, BI-RADS 4A, BI-RADS 4B, BI-RADS 4C y BI-RADS 5 y 6). El modelo produce cuatro logits por cada sesión. Para entrenar los modelos las mamografías se dividen cronológicamente en grupos de datos de entrenamiento y de validación. Todos los modelos se implementan utilizando el marco PyTorch DL.
Patrón de referencia	Definición de lesiones malignas: diagnóstico patológico en un plazo de dos años desde el momento en que la paciente acudió al hospital para la primera mamografía. Definición de las lesiones benignas: 1) el diagnóstico patológico de la misma lesión en un plazo de dos años fue benigno; y 2) las pacientes fueron objeto de seguimiento durante más de dos años, y la mamografía realizada más de dos años después del primer examen mamográfico indicó benignidad, sin diagnóstico patológico.
Flujo y tiempo	

Comparador	Para la evaluación clínica del modelo (2.ª parte) se compara la lectura radiológica frente a la lectura radiológica con ayuda del sistema de IA. Se evalúa la eficacia del modelo en la detección y el diagnóstico de mamografías controlando el rendimiento de 12 especialistas en radiología en diferentes condiciones de lectura. Los especialistas en radiología están blindados a cualquier información sobre los pacientes, incluidas imágenes previas e informes histopatológicos. La evaluación consiste en dos periodos de lectura separadas por cuatro semanas. A los especialistas en radiología se les informa que la tasa de malignidad en el grupo de datos evaluado es superior que la de la práctica clínica. Para cada caso, los especialistas en radiología emplean la clasificación BI-RADS (rango, 1-5), y etiquetan las lesiones sospechosas como benignas o malignas, y los pacientes normales sin lesiones se tienen en cuenta como negativos. Los especialistas en radiología puntúan cada caso en una escala de dificultad de 1-9 (9 representa la escala de dificultad más alta).
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto

DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	No		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	Sí		
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Poco claro		

¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	Poco claro		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Poco claro	

Autor: van Winkel *et al*. 2021	
Características del estudio	
Muestra de pacientes	Estudio de casos múltiples con múltiples lectores totalmente aleatorizado y cruzado, con 18 especialistas en radiología, que leen una serie de exámenes de BDT gran angular dos veces, con y sin apoyo de IA. El estudio incluye 360 casos recogidos entre mayo de 2011 y febrero de 2014: 110 casos de cáncer probados con biopsia, 104 casos benignos (probados con biopsia o al menos seis meses de seguimiento), y 146 casos negativos seleccionados aleatoriamente (al menos un año de seguimiento). Para la evaluación de los exámenes con y sin ayuda de IA se utilizan 240 casos (65 casos de cáncer, 65 casos benignos y 110 casos negativos). Los exámenes se adquieren utilizando ajustes de exposición estándar (Mammomat Inspiration) y están reconstruidos con el último algoritmo (EMPIRE) que también genera las imágenes mamográficas sintéticas.
Características de los pacientes y ámbito	
	Criterios de inclusión Mujeres a las que se realiza exámenes de BDT para diagnóstico y cribado, edad media 56,3 años. Exámenes bilaterales de dos vistas (CC y OML) de BDT.

	Criterios de exclusión Implantes mamarios. Calidad subóptima (juzgada por especialistas en radiología y radiógrafos experimentados). Datos de imagen o de verdad perdidos.
	Ámbito clínico Siete centros clínicos en EE. UU. representativos de mujeres sometidas a exámenes de BDT para cribado y diagnóstico.
Prueba índice	Sistema de IA Transpara™ 1.6.0. (ScreenPoint Medical BV), basado en redes neuronales convolucionales profundas, que automáticamente detecta lesiones sospechosas de cáncer de mama en mamografías 2D y BDT. Los resultados del sistema se muestran de dos maneras diferentes al especialista en radiología: 1) mediante una puntuación entre 1 y 10 que indica la probabilidad incremental de que un cáncer visible esté presente en la mamografía y 2) los hallazgos más sospechosos se marcan y se puntúan con el nivel de sospecha para cáncer (1-100). El sistema está validado para mamografías 2D en estudios previos con grupos de datos independientes. Ha sido entrenado y probado utilizando una base de datos propia que contiene sobre 1.000.000 de mamografías 2D y de imágenes BDT (sobre 20.000 con cáncer), adquiridas con máquinas de diferentes proveedores en una docena de instituciones de 10 países en Europa, América y Asia. Cada mamografía BDT se procesa por el sistema de IA. Los especialistas en radiología pueden utilizar simultáneamente el sistema de IA con o sin la mamografía sintética correspondiente y con o sin el apoyo de navegación interactiva que permite el acceso al plano de la BDT en donde el algoritmo de IA detecta anormalidades.
Patrón de referencia	Para cada caso, por mama, se basa en informes patológicos y de imagen (incluyen localización y caracterización radiológica del cáncer, localización de las lesiones benignas o confirmación de estatus normal) disponibles en formato electrónico y revisados por especialistas en radiología participantes en el proceso de selección de casos (no participan en el estudio observacional).

Flujo y tiempo	Se excluyen 120 casos (26 %), 75 en los que se señala el motivo y 45 en los que no. De los 75 casos excluidos, dos son implantes mamarios, 36 por objetos de reconstrucción, cuatro por pobre posicionamiento, uno por datos perdidos y 14 por seguimiento incompleto.
Comparador	Los exámenes se leen dos veces, con y sin ayuda de la IA, separadas por un periodo de limpieza de cuatro semanas. El orden de los casos y la disponibilidad del apoyo de la IA se asignaron aleatoriamente a cada especialista en radiología. Durante la evaluación de los casos con ayuda de la IA, se comprueban dos protocolos de lectura, la mitad de los especialistas en radiología leen los exámenes con acceso a las correspondientes mamografías sintéticas y apoyo interactivo a la navegación y la otra mitad los leen sin estas funciones. Los especialistas en radiología están cegados a la historia clínica del paciente y a cualquier otra información no visible en los exámenes de imágenes BDT incluidos. Para cada caso, los especialistas en radiología: marcan la localización 3D de los hallazgos en cada vista, asignan el nivel de sospecha a cada hallazgo, proveen una categoría BI-RADS por mama (1, 2, 3, 4a, 4b, 4c, 5). Todos los especialistas en radiología están certificados por la *American Board of Radiology*, cualificados para interpretar mamografías según la MQSA y activos en la lectura de exámenes DBT en la práctica clínica.
Notas	

Calidad metodológica			
Ítem	**Juicio de los autores**	**Riesgo de sesgo**	**Aplicabilidad**
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	Poco claro		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		

¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Poco claro		
Si se utilizó un umbral ¿se especificó previamente?	No		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).	No		
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	No		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	Sí		
		Alto	Alto

DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	Poco claro		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Poco claro	
Autor: Hsu *et al*. 2022			
Características del estudio			
Muestra de pacientes	Estudio de diagnóstico que utiliza datos clínicos, de imagen y de resultados de cáncer recogidos de un estudio observacional (Athena Breast Health Network) conducido a través de programas de cribado de mama. Exámenes de imagen mamaria notificados por el sistema de información radiológica: 49.244 mujeres y 184.935 imágenes. Mujeres incluidas después de las exclusiones realizadas 41.343 (723 con al menos un diagnóstico de cáncer de mama), 121.753 imágenes. Los exámenes calificados como VN se sometieron a un muestreo insuficiente, mientras que los calificados como FP, VP y FN se sometieron a un muestreo excesivo.		
Características de los pacientes y ámbito			
	Criterios de inclusión Mujeres que acuden a un centro ambulatorio de diagnóstico por imagen para someterse a una mamografía o ecografía mamaria (cribado o diagnóstico). El análisis se centra en mamografías de cribado 2D, obtenidas con un equipo de la casa Hologic.		

	Criterios de exclusión Imágenes que no pueden ser descargadas del PACS, que no disponen de un conjunto estándar de imágenes de cribado, o que pierden datos clínicos necesarios para ejecutar el modelo.
	Ámbito clínico Cinco centros médicos universitarios de California.
Prueba índice	CEM publicado en el DREAM Challenge, que incorpora predicciones de los 11 modelos con los mejores resultados, utilizando una población de cribado estadounidense independiente y diversa. Cada modelo se trata como una caja negra, no se realizó ninguna modificación en los algoritmos, que fueron entrenados con el conjunto de datos KPW antes de utilizarlos en el grupo de datos UCLA. Cada modelo genera una puntuación de confianza entre 0 y 1 que refleja la probabilidad de cáncer de cada lado de la mama. El CEM utilizó las puntuaciones de confianza de cada modelo, las volvió a ponderar y obtuvo una puntuación combinada. El CEM con sospecha del especialista en radiología (CEM+R) es el modelo de conjunto con la entrada añadida de la puntuación BI-RADS global proporcionada por el especialista en radiología intérprete original a nivel de examen. La puntuación BI-RADS se binarizó en baja (BI-RADS 1/2) y alta (BI-RADS 0/3/4/5) sospecha, y luego se añadió como variable independiente adicional.
Patrón de referencia	Diagnóstico histopatológico. Los exámenes se dividieron en cuatro grupos: • VN: exámenes de cribado anual con BI-RADS 1 y 2 sin diagnóstico de cáncer entre los exámenes. • FP: BI-RADS 0 y no diagnóstico de cáncer en 12 meses. • VP: BIRADS 0 y diagnóstico de cáncer en 12 meses. • FN: BI-RADS 1 y 2 diagnóstico de cáncer en 12 meses.
Flujo y tiempo	Se excluyeron 8.517 ecografías mamarias, 54.665 mamografías digitales con tomosíntesis y 7.901 mujeres sin exámenes de cribado 2D.
Comparador	El rendimiento del CEM y el CEM combinado con la evaluación del especialista en radiología (CEM+R) se compara con el diagnóstico realizado por especialista en radiología de carcinomas ductales *in situ* y cánceres invasivos diagnosticados en el plazo de un año desde el examen de cribado.

Notas			
Calidad metodológica			
Ítem	**Juicio de los autores**	**Riesgo de sesgo**	**Aplicabilidad**
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	Sí		
¿Evitó el estudio exclusiones inapropiadas?	Poco claro		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Poco claro		
		Bajo	Poco claro
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Bajo	Alto

DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	No		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Poco claro		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	Sí		
		Alto	Alto
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	No		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Alto	

Autor: Leibig *et al*. 2022	
Características del estudio	
Muestra de pacientes	Análisis retrospectivo que utiliza un conjunto de datos de 1.193.197 estudios de FFDM recogidos entre enero de 2007 y diciembre de 2020 de 453.104 mujeres. De seis centros de cribado se obtiene un grupo de datos de prueba interna para entrenamiento, divididos aleatoriamente en conjuntos de datos (mutuamente excluyentes) para entrenamiento, validación y prueba (no incluido en el entrenamiento y validación): 229.796 mujeres, 524.413 estudios de imagen (cánceres: 12.065, no incluidos cánceres omitidos o de intervalo). De dos centros de cribado se obtiene un grupo de datos de prueba externa de exámenes de cribado de cáncer de mama, no vistos anteriormente por el sistema de IA: 92.585 mujeres, 213.694 estudios de imagen (cánceres: 2.793, no incluidos cánceres omitidos o de intervalo). Los estudios sospechosos resueltos mediante consenso, incluidos los revalorados y los sometidos a biopsias, se sobremuestrean durante la recogida de datos. Para garantizar que los datos de prueba reflejan una población de cribado real, el enriquecimiento de cada conjunto de datos ocasionado por el sobremuestreo de casos de cáncer se maneja mediante técnica de ponderación.
Características de los pacientes y ámbito	
	Criterios de inclusión Mujeres asintomáticas, de entre 50-70 años y con una categoría de densidad mamaria B o C de acuerdo con la ACR, participantes cada dos años en el programa nacional de cribado de cáncer de mama alemán, en el que se utiliza un sistema de doble lectura realizada por dos especialistas en radiología cegados ante las decisiones del otro y en el que en el caso de que uno o los dos asignen a las imágenes mamográficas BI-RADS > 2 se realiza una reunión de consenso guiada por el especialista en radiología principal para reconciliar las diferencias en la interpretación. Las mamografías comprenden las cuatro vistas estándar, bilateral CC y OML (obtenidas utilizando dispositivos de diversos proveedores: Siemens, Hologic, Fugi, otros).
	Criterios de exclusión Imágenes de diagnóstico o de revaloración. Estudios de imagen normales sin seguimiento.

	Ámbito clínico Ocho centros de cribado en Alemania.
Prueba índice	Sistema de IA *in-house* que clasifica cáncer a nivel de estudio (entre 0 y 1). Está basado en una red neuronal convolucional profunda, entrenada con imágenes mamográficas utilizando etiquetas (obtenidas a partir de anotaciones de hallazgos radiológicos e información de biopsia) asociada a diferentes escalas solo con el propósito de entrenamiento. Para la IA como clasificación (triaje) se establecen dos umbrales para categorizar la derivación de decisiones: triaje normal, red de seguridad y derivación al especialista en radiología/radióloga. Los umbrales se representan como conjunto de dos puntos operativos establecidos para alcanzar la sensibilidad y especificidad deseadas en el conjunto de datos de validación: triaje normal (sensibilidad del algoritmo en el conjunto de datos de validación) + red de seguridad (sensibilidad del algoritmo en el conjunto de datos de validación). Para la IA como sistema independiente (*stand-alone*) el umbral de referencia establecido en el conjunto de datos de validación es igual a la sensibilidad del especialista en radiología (86 %).
Patrón de referencia	Los cánceres en el conjunto de datos se detectan por el cribado (se excluyen los cánceres omitidos o diagnosticados en el intervalo entre rondas de cribado). Se etiquetaron como positivos en base a confirmación histopatológica. Exámenes mamográficos normales: los derivados de mujeres con un periodo de seguimiento de cribado mínimo de 24 meses, no revaloradas (BI-RADS 1 o 2), o en el caso de hallazgos, estudio de seguimiento considerado negativo por doble lectura, consenso o revaloración negativa.
Flujo y tiempo	Se excluyen 314.869 y 140.221 estudios de imagen normales sin seguimiento en el grupo de datos de prueba interna y de datos de prueba externa, respectivamente. No se incluyeron los cánceres omitidos o diagnosticados en el intervalo entre rondas de cribado.

Comparador	Los resultados de cribado de un único especialista en radiología sin ayuda, basados en sus decisiones clínicas originales en el programa de cribado se comparan con los de un sistema de IA *stand-alone* y con un enfoque denominado de decisión-referencia (*decision-referral*) que combina la clasificación (triaje) normal y la detección de cáncer por medio de un sistema de alerta de red de seguridad. Las decisiones del especialista en radiología original son las registradas durante la práctica clínica sin apoyo de IA en el punto de lectura de pantalla antes de la conferencia de consenso o el arbitraje. Los análisis se limitan a los cánceres detectados por cribado y mamografías de seguimiento probadamente normales. En el enfoque decisión-referencia, el sistema de IA clasifica si un estudio es normal o sospechoso de cáncer y proporciona al mismo tiempo una indicación de su confianza en su clasificación, en base a dos umbrales. Los estudios sospechosos y los estudios para los que el algoritmo de IA no tenía confianza y requieren interpretación humana se mandan al especialista en radiología sin indicación de la clasificación del sistema de IA. Además, se evalúa una red de seguridad que se activa en los estudios que el sistema de IA considera sospechosos de cáncer.
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	No		
¿Evitó el estudio exclusiones inapropiadas?	No		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	Sí		
		Alto	Alto

DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Bajo	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Poco claro		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		

¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	No		
¿Se incluyó a todos los pacientes en el análisis?	No		
		Alto	

Autor: Romero-Martín *et al*. 2022	
Características del estudio	
Muestra de pacientes	Estudio cohortes que incluye 16.068 mamografías de cribado consecutivas e independientes (16.067 mujeres) recogidas retrospectivamente entre enero de 2015 y diciembre de 2016 del ensayo de cribado con tomosíntesis de Córdoba. Para cada mujer se recoge la edad, densidad mamaria, hallazgos histopatológicos de los procedimientos de biopsia y cáncer de intervalo diagnosticado.
Características de los pacientes y ámbito	
	Criterios de inclusión Mujeres participantes en el programa de cribado de cáncer de mama, edad 50-69 años. DM o BDT (dos vistas por cada mama) llevada a cabo con el dispositivo Selenia Dimensión (Hologic).
	Criterios de exclusión No se aplicaron criterios de exclusión.
	Ámbito clínico Una institución en Córdoba, España.

Prueba índice	Sistema de IA Transpara versión 1.7.0 (ScreenPoint Medical), comercialmente disponible. Utiliza algoritmos de aprendizaje profundo para analizar DM y BDT y detectar lesiones sospechosas de cáncer de mama. Los hallazgos más sospechosos detectados por el sistema son marcados en cada examen y se les asigna una puntuación entre 1 y 100, indicando la probabilidad incremental de que un cáncer visible esté presente en la mamografía. Cada examen de DM y BDT se procesa de forma independiente y da una puntuación IA para DM y otra para BDT. Un grupo de cuatro especialistas en radiología revisan los casos revalorados por la IA con el patrón de referencia para evaluar si la IA identifica correctamente las lesiones de cáncer. Solo se consideran hallazgos verdaderamente positivos de cáncer si la IA los localiza correctamente y los marca con la mayor puntuación posible. En caso de que no haya acuerdo entre especialistas en radiología, consenso.
Patrón de referencia	Hallazgos histopatológicos de los procedimientos de biopsia y cánceres de intervalo diagnosticados.
Flujo y tiempo	Se excluyeron los exámenes que no pudieron recuperarse del sistema de archivo y comunicación de imágenes (68 mujeres, cuatro revaloraciones no cancerosas y 64 VN).
Comparador	Se investiga si un sistema de IA por sí solo como herramienta de cribado para DM y BDT puede alcanzar sensibilidad similar con una aceptable tasa de revaloración en comparación con los cuatro escenarios originales con especialistas en radiología: lectura simple de DM, lectura doble de DM, lectura simple de BDT y lectura doble de BDT.
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	Sí		
¿Evitó el estudio exclusiones inapropiadas?	Sí		

¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	No		
		Bajo	Bajo
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Bajo	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo

DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	No		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Alto	
Autor: Sharma *et al*. 2021			
Características del estudio			
Muestra de pacientes	Cohorte histórica (de 10 años) consecutiva de pacientes, no enriquecida, en donde la edad de las pacientes, el intervalo de cribado y el método para la detección de cáncer es representativo de una población del mundo real. 304.360 casos de cribado compatible con FFDM (cuatro vistas) (275.900 casos (177.882 mujeres) eligibles en la muestra de 10 años, después de los casos excluidos).		
Características de los pacientes y ámbito			
	Criterios de inclusión Mujeres entre 50-70 años invitadas a participar en el programa de cribado de mama del R.U. (intervalo de cribado de tres años). También se incluyó una pequeña muestra de mujeres de entre 47-49 años y de entre 71-73 años. Mujeres de entre 45-65 años invitadas a participar en el programa de cribado de mama de Hungría (intervalo de cribado de dos años). También se incluyeron mujeres fuera del rango de edad que decidieron participar en el cribado oportunistamente.		

	Criterios de exclusión Casos que no eran de mujeres. Mujeres y casos seleccionados aleatoriamente para futuras investigaciones y desarrollos. Casos no procedentes del proveedor principal del centro, señalados como revaloraciones técnicas, que tienen una opinión incompleta del lector. Mujeres con historia comprobada de cáncer de mama.
	Ámbito clínico Siete centros europeos que representan 4 centros, 3 en el Reino Unido y 1 en Hungría. Los centros del Reino Unido participan en el programa de cribado de mama del servicio nacional de salud.
Prueba índice	Sistema de IA Mia ™ versión 2.0.1 (Kheiron Medical Technologies). La IA trabaja con casos de DICOM como input, analiza cuatro imágenes con dos vistas estándar de FFDM y genera una propuesta binaria de revalorar (más análisis debido a la sospecha de malignidad) o no revalorar (hasta la siguiente ronda de cribado). El sistema de IA utiliza umbrales predefinidos para revalorar o no. Los datos de los participantes en el estudio no se utilizaron en ningún aspecto de desarrollo del algoritmo de IA.

Patrón de referencia	La sensibilidad, la tasa de cánceres detectados y el valor predictivo positivo se define como positivos detectados en el cribado y cánceres posteriores a los 3 años. Los positivos detectados en el cribado fueron casos de cribado correctamente identificados por el flujo de trabajo histórico de doble lectura, con tumores malignos con patología confirmada por citología, biopsia y/o histología a los 180 días del examen de cribado. La especificidad se define como cualquier caso de cribado con evidencia de un resultado de seguimiento negativo que incluye lectura de mamografías de al menos 1.035 días (dos meses menos que los tres años de la ronda de cribado) después de la fecha del cribado original, sin ninguna prueba de malignidad en el medio. La tasa de recuperación, la tasa de cánceres detectados y la tasa de revaloraciones se calcularon sobre toda la población, que incluía los casos positivos confirmados, los negativos confirmados y los no confirmados. Los cánceres posteriores a los tres años se definen como un caso de cribado con un cáncer con patología probada surgido en los 1.095 días siguientes a la fecha original de cribado y se alinean con la definición de cánceres de intervalo para los programas de intervalo de cribado de tres años, como en el R.U. El intervalo de cribado de dos años seguido en Hungría significa que todos los cánceres de intervalo dentro del intervalo de cribado de dos y los cánceres adicionales detectados en la siguiente ronda de cribado, también se incluyen como casos de «cáncer posterior de tres años».
Flujo y tiempo	Se excluyen 28.460 (9,59 %) de los casos de acuerdo con los criterios de exclusión señalados.
Comparador	Comparar el rendimiento de un sistema de IA *stand-alone* con el histórico primer lector humano. Simular el rendimiento de la doble lectura utilizando la IA como segundo lector independiente comparado con el histórico doble lector humano. En la doble lectura sin IA la opinión del primer lector se hace de forma aislada, y el segundo lector tiene acceso, a su criterio, a la opinión del primero y en caso de acuerdo deciden revalorar o no. En caso de desacuerdo, un arbitraje, realizado por un especialista en radiología o grupo de especialistas en radiología, toman la decisión definitiva. En la doble lectura con IA, cuando la IA y el primer lector están de acuerdo, se toma una decisión de revalorar o no. En caso de desacuerdo, si está disponible, se utiliza el dictamen histórico de arbitraje, o se elige la opinión del segundo lector.
Notas	

Calidad metodológica			
Ítem	**Juicio de los autores**	**Riesgo de sesgo**	**Aplicabilidad**
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	Sí		
¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	No		
		Bajo	Bajo
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	No		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Alto	Alto

DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		
¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	No		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Alto	
Autor: Larsen *et al*. 2022			
Características del estudio			
Muestra de pacientes	Estudio basado en datos retrospectivos de cuatro unidades de cribado del programa de cribado de mama noruego (programa poblacional de cribado). 47.877 mujeres cribadas, 122.969 exámenes de cribado (957 cánceres).		
Características de los pacientes y ámbito			

	Criterios de inclusión Mujeres entre 50 y 69 años a las que el cribado de mama noruego ofrece cribado mamográfico de dos vistas bienal. Mamografías digitales obtenidas con MAMMOMAT (Siemens Healthcare).
	Criterios de exclusión Mamografías con problemas técnicos en el proceso de extracción y de puesta de seudónimos. Revaloraciones debidas a síntomas autodeclarados en el cribado. Revaloraciones debidas a mamografías técnicamente inadecuadas.
	Ámbito clínico Unidades de cribado del cribado de mama noruego.
Prueba índice	Sistema de IA Transpara 1.7.0 (ScreenPoint Medical), comercialmente disponible, utiliza redes neuronales convolucionales para analizar mamografías y está entrenado con mamografías de diferentes programas de cribado y varios proveedores. Provee una puntuación para cada vista de cada mama. Se utiliza la puntuación más alta de todas las vistas para asignar una puntuación global a nivel de examen (entre 1 y 10) basada en una puntuación bruta. Se explora el desempeño del sistema de IA como una herramienta de decisión binaria con tres diferentes umbrales para seleccionar si los exámenes son sospechosos o no. Umbral 1: examen con una puntuación bruta > 9.00 (puntuación global de 10) se define como «seleccionado» por el sistema de IA, con una puntuación < 10 «no seleccionado». Umbral 2: tasa de selección = 8,8 % (tasa de consenso) (puntuación bruta > 9,13), utilizado para explorar el desempeño de la IA cuando el número de exámenes seleccionados por el sistema como sospechoso es similar al número de exámenes seleccionados por los dos especialistas en radiología. Umbral 3: tasa de selección = 5,8 % (puntuación bruta > 9,43) que es la tasa individual media de interpretaciones positivas observadas por los especialistas en radiología en la muestra de estudio.

Patrón de referencia	Examen negativo: si las mamografías tienen una evaluación negativa por ambos especialistas en radiología, después de consenso o tienen una revaloración con resultado negativo. Revaloración: exámenes de cribado que resultan en más evaluaciones debido a hallazgos mamográficos anormales. Cáncer detectado por el cribado: cáncer de mama diagnosticado después de una revaloración y en los seis meses siguientes al examen de cribado. Cáncer de intervalo: cáncer de mama diagnosticado en los 24 meses siguientes a un examen de cribado negativo o a los 6-24 meses siguientes a una revaloración con un resultado negativo.
Flujo y tiempo	8.740 exámenes, 414 mujeres excluidas debido a mamografías con problemas técnicos en el proceso de extracción y de puesta de seudónimos. 330 exámenes, 318 mujeres excluidas por revaloraciones debidas a síntomas autodeclarados en el cribado. 156 exámenes, 156 mujeres excluidas por revaloraciones debidas a mamografías técnicamente inadecuadas.
Comparador	Se compara el desempeño de un sistema de IA comercialmente disponible con la doble lectura, en la que dos especialistas en radiología independientes interpretan las mamografías y les asignan una puntuación de interpretación entre 1 y 5 para indicar la sospecha de malignidad (1 negativo para malignidad, 2 probablemente benigno, 3 sospecha de malignidad intermedia, 4 probablemente maligno y 5 sospecha alta de malignidad). Si la puntuación de interpretación es 2 o mayor por cada especialista en radiología, un consenso de al menos dos especialistas en radiología determina si revalorar a la mujer o no.
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	Sí		
¿Evitó el estudio exclusiones inapropiadas?	Sí		

¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?	No		
		Bajo	Bajo
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Bajo	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		

¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?	No		
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice).	No		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Alto	
Autor: Lauritzen *et al*. 2022			
Características del estudio			
Muestra de pacientes	Estudio retrospectivo que analiza datos de exámenes mamográficos consecutivamente recogidos en enero de 2014 y diciembre de 2015. Se utilizaron dos muestras de cribado. La muestra de desarrollo incluye 54.997 mujeres cribadas consecutivamente, previamente reportadas. 53.951 mujeres después de excluir las vistas perdidas y los datos de imagen incompletos (1.046 mujeres). La muestra de prueba incluye 118.039 mujeres consecutivamente cribadas. 114.421 mujeres después de excluir las vistas perdidas y los datos de imagen incompletos (3.618 mujeres).		

Características de los pacientes y ámbito	
	Criterios de inclusión Mujeres asintomáticas de 50-69 años de la Región Capital de Dinamarca a las que se ofrece cribado bienal del cáncer de mama mediante sistemas Mammomat Inspiration (Siemens Healthineers). Exámenes mamográficos de cribado, mamografías digitales, de cuatro vistas (vistas CC y OML de cada mama), leídas independientemente por dos especialistas en radiología (si desacuerdo con respecto a las revaloraciones, consenso con un tercer especialista en radiología).
	Criterios de exclusión Vistas perdidas o datos de imagen incompletos.
	Ámbito clínico Programa de cribado de cáncer de mama en la Región Capital de Dinamarca.

Prueba índice	Sistema de IA Transpara versión 1.7.0 (Screen Point Medical). Utiliza redes neuronales convolucionales profundas y está entrenado en más de un millón de mamografías de distintos sitios de Europa y EE.UU., para detectar lesiones sospechosas de cáncer de mama en FFDM capturadas en máquinas de diferentes proveedores. El análisis de resultados comprende una puntuación de examen entre 0 y 10, indicando el riesgo de presencia de posible cáncer visible, calibrada de tal manera que el 10 % de una población entra dentro de cada una de las 10 categorías. Transpara se utiliza con la configuración predeterminada y los datos del estudio son completamente independientes de los utilizados en el desarrollo del sistema de IA. Se utilizan dos umbrales para categorizar qué mamografías son normales, de riesgo moderado y sospechosas. El umbral de exclusión de 5 significa que aproximadamente el 50 % de mamografías se categorizan como normales y el UR de 9,989, que se deriva aplicando el sistema de IA a la muestra de desarrollo, se ajusta de forma que el número de cánceres no detectados en el cribado (por IA) sea igual al número de exámenes de cribado sospechosos diagnosticados posteriormente como cáncer de intervalo. En consecuencia, los cribados radiológicos y los basados en IA se igualan en sensibilidad en la muestra de desarrollo: • Una nota de examen < 5 se categoriza la mamografía como normal. • Una nota de examen ≥ 5 pero ≤ que el UR se categoriza como de riesgo moderado, por lo que la leen los dos especialistas en radiología (las decisiones de revaloración del especialista en radiología se extraen de los informes de cribado originales). • Una nota de examen > que el UR se categoriza como sospechosa, no fue leída por los especialistas en radiología y la mujer se revalora directamente.

Patrón de referencia	Las mujeres con resultados positivos de cribado (hallazgos sospechosos) son invitadas a la realización de una prueba triple que consiste en: examen clínico, examen con mamografía y ecografía, biopsia, si es relevante, y/o más imágenes de la mama (resonancia magnética). Mujeres sin hallazgos sospechosos (resultados negativos de cribado) se les vuelve a invitar a participar en el cribado dos años después. Diagnóstico en base a biopsias y resultados. Cáncer detectado mediante el cribado: mujeres con cribado positivo diagnosticadas de cáncer de mama o carcinoma ductal *in situ* en los seis meses siguientes al cribado. FP: mujeres con cribado positivo, pero sin cáncer detectado mediante cribado. Cánceres de intervalo: cánceres en mujeres con un cribado o revaloración negativos diagnosticadas en los 24 meses siguientes al cribado (descubiertos fuera del cribado y confirmados con la prueba triple). Cánceres a largo plazo: cánceres de mama diagnosticados a los 2-5 años después del cribado.
Flujo y tiempo	1.046 mujeres excluidas en la muestra de desarrollo debido a vistas perdidas y los datos de imagen incompletos. 3.618 (3 %) mujeres excluidas en la muestra de prueba debido a vistas perdidas y los datos de imagen incompletos. Cabe suponer que, en el caso del cribado basado en IA aplicado *in situ* en clínicas, no se perdería ningún dato de imagen.
Comparador	Se analiza realizar el cribado solo con el sistema de IA, en donde este sustituye a los dos lectores y en comparación con el cribado en el que dos especialistas en radiología leen todas las mamografías sin interacción con ningún sistema de IA.
Notas	

Calidad metodológica

Ítem	Juicio de los autores	Riesgo de sesgo	Aplicabilidad
DOMINIO 1: Selección de participantes			
¿Se reclutó una muestra consecutiva o aleatorizada de pacientes?	Sí		

¿Evitó el estudio exclusiones inapropiadas?	Sí		
¿Las mujeres y mamografías incluidas en el estudio fueron independientes de aquellas utilizadas en el entrenamiento del algoritmo de IA?	Sí		
¿Existe preocupación de que los pacientes incluidos no coincidan con los de la pregunta de la revisión?			
		Bajo	Bajo
DOMINIO 2: Prueba índice			
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados del patrón de referencia?	Sí		
¿Se interpretaron los resultados de la prueba índice sin conocer los resultados de cualquier otra prueba índice?	Sí		
Si se utilizó un umbral ¿se especificó previamente?	Sí		
Donde los lectores humanos fueron parte de la prueba ¿tomaron sus decisiones en un contexto de práctica clínica? (es decir, evitando el efecto laboratorio).			
¿Existe preocupación de que la(s) prueba(s) índice o el comparador, su realización o interpretación difieran de la pregunta de revisión?	Sí		
		Bajo	Alto
DOMINIO 3: Patrón de referencia			
¿Es probable que el patrón de referencia clasifique correctamente la condición diana?	Sí		
¿Se interpretaron los resultados del patrón de referencia sin conocer los resultados de la prueba índice?	Sí		

¿Existe preocupación de que la condición diana definida por el patrón de referencia no coincida con la pregunta de revisión?			
		Bajo	Bajo
DOMINIO 4: Flujo y tiempo			
¿Recibieron todos los pacientes un patrón de referencia?	Sí		
¿Evitó el estudio elegir el patrón de referencia basándose en los resultados de una sola de las pruebas índice? (Todos los estudios tendrán necesariamente una verificación diferencial, porque no todas las mujeres pueden o deben someterse a una biopsia. Aquí se está midiendo si decidir que se reciba un patrón de referencia basado en los resultados de una sola de las pruebas índice.)	No		
¿Se incluyó a todos los pacientes en el análisis?	Sí		
		Alto	

Anexo VI.3. Estudios excluidos

VI.3.1. Revisiones sistemáticas

Autor	Título	Motivo
Taylor-Phillips, S., 2022	UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening.	No RS. Narrativa
Rautela, K., 2022	A Systematic Review on Breast Cancer Detection Using Deep Learning Techniques.	No responde a las preguntas de investigación
Malliori, A., 2022	Breast cancer detection using machine learning in digital mammography and breast tomosynthesis: A systematic review.	No análisis de la IA dentro de PDPCM
Lamb, L.R., 2022	Artificial Intelligence (AI) for Screening Mammography, From the AJR Special Series on AI Applications.	No RS. Narrativa
Jairam, M.P., 2022	A review of artificial intelligence in mammography.	No RS. Narrativa
Hickman, S.E., 2022	Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis.	No análisis de la IA dentro de PDPCM
Anderson, A.W., 2022	Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review.	No análisis de la IA dentro de PDPCM
Batchu, S., 2021	A Review of Applications of Machine Learning in Mammography and Future Challenges.	No RS. Narrativa
Masud, R., 2019	Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review.	No RS. De alcance
Houssami, N., 2019	Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice.	No RS. De alcance
Houssami, N., 2009	Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening.	No RS. Narrativa
Boyer, B., 2009	CAD in questions/answers Review of the literature.	No RS

VI.3.2. Estudios primarios

Autor	Título	Motivo
Yoon, J., 2022	AI-CAD for differentiating lesions presenting as calcifications only on mammography: outcome analysis incorporating the ACR BI-RADS descriptors for calcifications	Sistema de IA para la detección de subtipos de cáncer
Yala, A., 2022	Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model	El Sistema de IA se utiliza para predecir el riesgo futuro de cáncer
Wanders, A.J.T., 2022	Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms	Tipo de estudio: caso-control. Validación cruzada. Solo se incluyen pacientes con cánceres de intervalo
Wan, Y., 2022	Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography	Sistema de IA para la detección de subtipos de cáncer
Shoshan, Y., 2022	Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis	Los datos utilizados para generalizar el modelo son una muestra de los datos utilizados para su entrenamiento, validación y prueba
Pinto, M.C., 2021	Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with Single-View Wide-Angle Digital Breast Tomosynthesis	BDT de una sola vista. No BDT de dos vistas
Park, G.U., 2022	Retrospective Review of Missed Cancer Detection and Its Mammography Findings with Artificial-Intelligence-Based, Computer-Aided Diagnosis	Solo se incluyen imágenes con cáncer probado con biopsia
Lehman, C.D., 2022	Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening	El Sistema de IA se utiliza para predecir el riesgo futuro de cáncer
Lang, K., 2021	Can artificial intelligence reduce the interval cancer rate in mammography screening?	Solo se incluyen imágenes con cáncer de intervalo
Yirgin, I.K., 2022	Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis	Evalúa el desarrollo de un sistema de IA

Autor	Título	Motivo
Kim, Y.S., 2022	Use of Artificial Intelligence for Reducing Unnecessary Recalls at Screening Mammography: A Simulation Study	Solo se incluyen mujeres revaloradas después del cribado mamográfico
Kerschke, 2022	Using deep learning to assist readers during the arbitration process: a lesion-based retrospective evaluation of breast cancer screening performance	Solo se incluyen mujeres revaloradas después del cribado mamográfico
He, Z., 2022	Can a Computer-Aided Mass Diagnosis Model Based on Perceptive Features Learned From Quantitative Mammography Radiology Reports Improve Junior Radiologists' Diagnosis Performance? An Observer Study	Conjunto de pruebas de validaciones divididas
Graewingholt, 2021	Retrospective analysis of the effect on interval cancer rate of adding an artificial intelligence algorithm to the reading process for two-dimensional full-field digital mammography	Solo se incluyen imágenes con cáncer de intervalo
Gastounioti, A., 2022	External Validation of a Mammography-Derived AI-Based Risk Model in a U.S. Breast Cancer Screening Cohort of White and Black Women	Tipo de estudio: caso-control. El Sistema de IA se utiliza para predecir el riesgo futuro de cáncer
Do, Y.A., 2021	Diagnostic Performance of Artificial Intelligence-Based Computer-Aided Diagnosis for Breast Microcalcification on Mammography	Sistema de IA para la detección de subtipos de cáncer
Dahlblom, V., 2021	Artificial Intelligence Detection of Missed Cancers at Digital Mammography That Were Detected at Digital Breast Tomosynthesis	Tipo de imágenes: DM+BDT
Chang, Y.W., 2022	Artificial Intelligence for Breast Cancer Screening in Mammography (AI-STREAM): A Prospective Multicenter Study Design in Korea Using AI-Based CADe/x	Tipo de estudio: protocolo
Buda, M., 2021	A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images	Conjunto de pruebas de validaciones divididas
Bing, D., 2022	AI-based prevention of interval cancers in a national mammography screening program	Solo se incluyen imágenes con cáncer de intervalo

VI.3.3. Estudios de evaluación económica

Autor	Título	Motivo
Mayo, R.C., 2019	Impact of Artificial Intelligence on Women's Imaging: Cost-Benefit Analysis	No es un estudio de EE

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco