XXVI Curso de actualización
Medicina Interna
2026

UNIVERSIDAD DE ANTIOQUIA
Facultad de Medicina

Manual de lectura crítica de artículos científicos con inteligencia artificial en medicina interna

Alejandro Hernández Arango
Profesor, Médico Internista del Departamento de Medicina Interna, Facultad de Medicina, Universidad de Antioquia,
Magister Cum Laude en Telesalud, Universidad de Antioquia,
Líder de Informática Médica GHIPS, Hospital Alma Mater, Universidad de Antioquia.

Caso clínico inicial relacionado

Usted es uno de los internistas del hospital y un día, le solicitan a su equipo clínico evaluar la posibilidad de integrar un software como dispositivo médico con inteligencia artificial (IA) para la tamización de cáncer de mama dentro de la historia clínica del cual, los fabricantes internacionales aducen que tiene "buena evidencia científica basada en ensayos clínicos" y proporcionan el estudio MASAI (1,2) el cual evaluó la no inferioridad de la tamización de cáncer de mama con mamografía asistida por IA (Grupo Intervención: IA + Humano) frente a la lectura doble estándar realizada únicamente por radiólogos (grupo control). Los resultados finales demostraron que el grupo asistido por IA logró una tasa de detección de cáncer superior, identificando 6,4 casos por cada 1.000 participantes (338 cánceres totales), en comparación con los 5,0 por 1.000 (262 cánceres) del grupo de control. Esta diferencia representa un aumento significativo del 29 % en la detección de cánceres.

En cuanto a la precisión diagnóstica, la IA mejoró el valor predictivo positivo (VPP) de las llamadas a revisión al 30,5 %, al superar el 25,5 % del grupo humano. Crucialmente, tras completar el seguimiento de dos años, se determinó que la sensibilidad real del sistema fue del 80,5 % (no del 100 %), el cual superó significativamente la sensibilidad del 73,8 % del grupo de control y permitió que el grupo con asistencia de IA identificara 270 cánceres invasivos (frente a los 217 del grupo humano), con lo que logró además una reducción del 44,2 % en la carga de trabajo de lectura y una disminución del 12 % en la tasa de cánceres de intervalo (falsos negativos a largo plazo).

Con esta evidencia, ¿Considera usted que sería útil para su clínica implementar un programa de tamización de cáncer de mama apoyado en este software con Inteligencia Artificial?

Tabla 1. Métricas del estudio MASAI (1,2)

IA - Mamografía (Resultados Finales)	Cálculos y Resultados
Total, pacientes	53.043
Positivo (Total cánceres)	420 (338 detectados + 82 de intervalo)
Negativo (Total sin cáncer)	52.623 (Total - Positivos Reales)
Verdadero positivo (TP)	338 (Cánceres detectados en el cribado)
Falso negativo (FN)	82 (Cánceres de intervalo)*
Sensibilidad (recall) = TP / (TP + FN)	338 / 420 = 80,5 %
Falso positivo (FP)	771 (Total recalls - TP - 1 intervalo recordado)
Verdadero negativo (TN)	51.852

Continúa en la siguiente página.

Tabla 1. Métricas del estudio MASAI (1,2) (Continuación)

Especificidad = TN / (FP + TN)	51.852 / (771 + 51.852) = 98,5 %
(Prevalencia) Total positivos de la población	420 / 53.043 ≈ 0,79 % (Prevalencia real confirmada)
Valor predictivo positivo (precisión)	30,5 % (Medido como cánceres detectados / total recalls)
Valor predictivo negativo = TN / (FN + TN)	51.852 / (82 + 51.852) ≈ 99,84 %
Accuracy = (TP + TN) / población	(338 + 51.852) / 53.043 ≈ 98,39 %
F1 score = 2 × (Precisión × recall) / (Precisión + recall)	2 × (0,305 × 0,805) / (0,305 + 0,805) = 0,491 / 1,11 ≈ 44,2 %

Introducción

La Inteligencia artificial es una gran promesa con potenciales aplicaciones en salud. En Medicina Interna específicamente el razonamiento clínico integra diferentes aspectos complejos que podrían potenciarse con uso de estas nuevas herramientas tecnológicas (3) en el 2026 en el momento de la escritura de este artículo la FDA contiene una lista de 1.317 *software as medical device* (SaMD) con inteligencia Artificial en ruta de aprobación (3) por esto, es posible que la tendencia de la toma de decisiones integrada a SaMD este cada vez más presente además de la Medicina Basada en la Evidencia en la vida del Internista.

Los artículos científicos que dan soporte a estas herramientas presentan desafíos únicos de interpretación que los hacen especialmente complejos dada la creciente necesidad de **validar el impacto clínico real** más allá de la precisión técnica o *in silico*. Este manual pretende hacer un resumen práctico de los estándares internacionales para que el internista pueda discernir si una herramienta de IA es confiable y aplicable a sus pacientes y su contexto.

Conceptos básicos de inteligencia artificial y analítica para el clínico

En esta sección haremos un glosario de términos que permita al lector acercarse a una lectura crítica con fundamentos básicos que permitan estructurar un concepto más claro sobre este tipo de enfoque (4):

- **Aprendizaje de máquina:** un campo de estudio que permite que las computadoras aprendan sin programación explícitamente.
- **Aprendizaje supervisado:** implica de alguna manera modelar la relación entre las características de los datos y alguna etiqueta asociada a estos; una vez que se determina el modelo, se puede usar para asociar etiquetas a datos nuevos y desconocidos. El aprendizaje supervisado se subdivide en clasificación y regresión: En clasificación se tienen etiquetas discretas, mientras que en la regresión, las etiquetas son cantidades continuas.
- **Aprendizaje no supervisado:** implica modelos que describen datos sin etiqueta conocida. Un caso común de aprendizaje no supervisado es el de agrupamiento (clustering) en el que los datos se asignan automáticamente a un cierto número de grupos discretos.
- **Red neuronal:** modelo matemático inspirado en el sistema nervioso central que se compone de unidades llamadas perceptrones que están organizadas en capas, la forma en la cual se organizan estas capas se denomina arquitectura de la red, y en conjunto con la cantidad de capas se puede denominar redes neuronales profundas de tipo convolucional, por ejemplo, las cuales han revolucionado el campo de la clasificación de imágenes dando un gran avance en visión de máquina.

- **Procesamiento natural del lenguaje:** tipos de modelo que se enfocan en trabajar todo lo relacionado con el procesamiento de texto libre como dato no estructurado
- **Instancia:** A una sola fila de datos se le llama instancia. También se le conoce como una observación del dominio.
- **Característica (Feature):** A una sola columna de datos se le llama característica. Es un componente de una observación y también se denomina atributo de una instancia de datos (la característica se suele asociar con el atributo y su valor, aunque la mayoría de las veces se usa atributo y característica indistintamente). Algunas características pueden ser entradas a un modelo (predictores) y otras pueden ser salidas o las características a predecir (también llamadas labels).
- **Datos de entrenamiento:** Conjunto de datos que introducimos a nuestro algoritmo para entrenar nuestro modelo.
- **Datos de prueba:** Conjunto de datos que utilizamos para validar la precisión de nuestro modelo pero que no se utiliza para entrenarlo. Puede llamarse conjunto de datos de validación. conjunto de datos donde se ajustan los hiperparámetros del modelo
- **Parámetros del modelo:** Son aquellos pesos que se generan dentro del modelo utilizado para realizar la predicción
- **Hiperparámetros:** Es un parámetro de un algoritmo de aprendizaje (no del modelo). Como tal, no se ve afectado por el algoritmo de aprendizaje en sí; debe establecerse antes al entrenamiento y permanece constante durante el entrenamiento.
- **Métrica:** Medida cuantitativa usada para evaluar el rendimiento del algoritmo.
- **Analítica de datos** es la aplicación de técnicas de procesamiento de datos contra grandes bases de datos en tiempo real, es descriptiva cuando se usan datos de tiempo pasado, predictiva cuando se genera predicciones sobre esos datos pasados, y prescriptiva cuando tiene la capacidad de simular escenarios futuros para ilustrar las consecuencias de una decisión.
- **Sistema de soporte a las decisiones clínicas:** Los sistemas de soporte a las decisiones clínicas (SSDC) son herramientas clínicas (no necesariamente digitales por ejemplo el nomograma de anticoagulación con heparina) diseñadas para que el médico, en un contexto adecuado, pueda tomar una decisión soportada; un sistema que le da elementos que le permita apoyarse en su razonamiento y toma de decisiones. Pueden ser sistemas expertos o basados en inteligencia artificial sin ser excluyentes.

Interpretación de estudios que usan inteligencia artificial a la luz de la medicina basada en la evidencia

La integración de la IA en la Medicina Interna debe seguir un rigor científico equivalente al de los fármacos. Desde el punto de vista de la investigación, un sistema de soporte a decisiones clínicas (CDSS) basado en IA debe transitar por cinco fases evolutivas, cada una con estándares de calidad específicos publicados en **DECIDE-AI** (5) el primer paso es identificar la fase de desarrollo en la cual está la IA que se está evaluando para utilizar la herramienta de calidad Adecuada (**Figura 1**):

Desarrollo preclínico (***In Silico***): Es la etapa inicial de creación y entrenamiento del algoritmo con datos históricos. El estándar **PROBAST+AI** (6) (en su componente de desarrollo) es fundamental en esta fase, ya que permite evaluar la calidad metodológica del proceso de construcción o "fabricación" del modelo. El internista debe verificar aquí si hubo transparencia en las fuentes de datos y si el diseño del algoritmo evitó errores fundamentales como la división correcta de los datos en entrenamiento ajuste y validación interna antes de pasar a pruebas clínicas. Generalmente son estudios observacionales retrospectivos.

Validación *offline:* En esta fase se pone a prueba el modelo con datos que no fueron usados en el entrenamiento. Normalmente **son estudios prospectivos de validación externa** y es importante entender que su naturaleza está relacionada a que el clínico o el paciente no usan el sistema, sino que la IA corre de forma silenciosa y los investigadores evalúan sus resultados. Aquí, **PROBAST+AI** (6) (componente de evaluación) se utiliza para medir el riesgo de sesgo en el rendimiento reportado, mientras que los estándares **TRIPOD-AI** (7) para estudios de pronóstico y **STARD-AI** (8) para diagnóstico dictan cómo deben reportarse los resultados técnicos para asegurar que la precisión no sea un “espejismo estadístico”.

Seguridad y utilidad a pequeña escala: Es la transición al entorno clínico real. El estándar **DECIDE-AI** (5) guía estos estudios prospectivos o pseudoexperimentales, enfocados no solo en la precisión, sino en la utilidad clínica inicial y la seguridad del paciente en manos de usuarios humanos (médicos

internistas reales en condiciones de práctica).

Seguridad y efectividad a gran escala (RCT): Es el estándar de oro para demostrar impacto en salud. Aquí se utiliza el **SPIRIT-AI** (9) para los protocolos de RCT y los informes de resultados de ensayos clínicos el **CONSORT-AI** (10) para asegurar que el uso de la IA realmente mejora desenlaces clínicos (ej. reducción de mortalidad por sepsis) frente al manejo convencional.

Vigilancia *post-marketing:* Una vez implementado, el algoritmo debe ser monitoreado continuamente para detectar "deriva del modelo" (pérdida de precisión con el tiempo) o efectos adversos imprevistos en diferentes poblaciones. Aquí el Marco **RE-AIM** (11) en la Implementación de IA juega un papel importante, además la **guía de IA la OMS** (12) para el uso trasparente y evaluación continua a través de los *model cards.*

Figura 1. Fases de desarrollo de un sistema de soporte a las decisiones clínicas basado en inteligencia artificial y sus guías de validación en *Equator-Network.*

Adaptado de *Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022 May;28(5):924–33. (5)*

El siguiente paso es aplicar el conjunto adecuado de preguntas sistemáticamente al reporte del estudio para tener una guía clara y crítica sobre la calidad de la evidencia (ver **Tabla 3**).

Revisiones sistemáticas potenciadas por IA: requieren reportes claros y modelos especializados

El rápido y constante crecimiento de la literatura científica ha impulsado la integración de la inteligencia artificial (IA) en la síntesis de evidencia y las revisiones sistemáticas de la literatura (SLR). Existen herramientas innovadoras como *OpenScholar* demuestran el enorme potencial de los grandes modelos de lenguaje (LLM) mejorados por retornos aumentados por recuperación (RAG) para asistir a los investigadores en la búsqueda y síntesis de literatura, superando a los modelos tradicionales al reducir drásticamente las alucinaciones y generar respuestas respaldadas con citas precisas extraídas de millones de artículos (13). Por lo que la primera recomendación es evitar modelos de propósito general para asistir la escritura de revisiones sistemáticas

Si lo que se desea es revisar artículos sobre IA en salud, la comunidad científica está desarrollando pautas de presentación de informes especializadas basadas en la declaración PRISMA (14) como extensión PRISMA-AI (15), diseñada para estandarizar las revisiones sistemáticas y metaanálisis que evalúan intervenciones basadas en IA como objeto de estudio en la atención médica.

De forma complementaria, se ha propuesto la lista de verificación PRISMA-trAIce (16), un marco agnóstico a la disciplina diseñado específicamente para documentar de manera transparente cuando la IA se utiliza como una herramienta metodológica dentro del propio proceso de revisión (por ejemplo, para la selección automatizada de estudios o la extracción de datos) agregando detalles cruciales como la identificación de la herramienta, la ingeniería de instrucciones (*prompts*), la interacción humano-IA y la evaluación del rendimiento al reporte.

Algunos sesgos especiales para tener en cuenta

La herramienta **PROBAST+AI** (6) de la cual la UdeA participó por Latinoamérica en su construcción, organiza formalmente el riesgo de sesgo en estudios que usan IA en salud en cuatro dominios clave que deben ser evaluados en todo modelo de predicción; el de **participantes**, relacionado a fuentes de datos inapropiadas o reclutamiento sesgado. El dominio de los **predictores** si los datos de entrada no se definieron o evaluaron de forma consistente entre los grupos de riesgo. El de **resultado** (***outcome***) cuando la determinación del resultado fue influenciada por el conocimiento de los predictores y; finalmente los sesgos relacionados con el **análisis** cuando esta Sesgado por un tamaño de muestra insuficiente, manejo inadecuado de datos faltantes o mala calibración del modelo. Existen además otros sesgos o definiciones de estos a conocer y tener en cuenta en la lectura crítica:

- **Sesgo de selección (*selection bias*):** Mencionado en JAMA Guide y PROBAST+AI. Ocurre cuando la población utilizada para entrenar el modelo no representa adecuadamente a la población donde se aplicará, o cuando los criterios de inclusión/exclusión en un ensayo clínico (mencionado en CONSORT-AI) favorecen sistemáticamente a ciertos grupos.
- **Pobreza de datos (*health data poverty*):** Citado en TRIPOD+AI, se refiere a la falta de representación de ciertos grupos demográficos en los conjuntos de datos, lo que genera modelos que funcionan peor para minorías.
- **Sesgo del estándar de referencia (*reference standard bias*):** Destacado en JAMA Guide y STARD-AI. Si el "Gold Standard" (la verdad absoluta) utilizado para entrenar la IA tiene errores humanos o inconsistencias, el modelo aprenderá y replicará esos mismos sesgos.
- **Fuga de datos (*data leakage*):** Aunque es un error técnico, actúa como un sesgo de optimismo. Ocurre cuando la información del conjunto de prueba "se filtra" al de entrenamiento, y hace que el modelo parezca mucho más preciso de lo que realmente es (JAMA Guide).
- **Sesgo algorítmico (*algorithmic bias*):** Mencionado en STARD-AI y DECIDE-AI como una preocupación central de equidad. Se refiere a errores sistemáticos en los que el algoritmo produce resultados diferentes para grupos basados en características como raza, género o nivel socioeconómico.
- **Sobreajuste (*overfitting*):** Considerado en PROBAST+AI y TRIPOD+AI como un sesgo que afecta la generalización; el modelo aprende "ruido" de los datos de entrenamiento en lugar de patrones reales, y falla al ser aplicado a nuevos pacientes.
- **Sesgo de automatización (*automation bias*):** Mencionado en DECIDE-AI y CONSORT-AI. Es la tendencia de los humanos a confiar excesivamente en las sugerencias de un sistema automatizado, incluso cuando contradicen su propio juicio clínico o la evidencia clara.
- **Sesgo de confirmación:** Los usuarios pueden utilizar los resultados de la IA para confirmar sus propias sospechas preexistentes, ignorando señales de advertencia (discutido en la evaluación de factores humanos de DECIDE-AI).
- **Sesgo de optimismo:** Los autores suelen reportar solo las métricas de desempeño más favorables (como el AUC más alto), y omiten medidas de error o desempeño en subgrupos específicos (TRIPOD+AI).

Marco normativo de los sistemas de soporte a las decisiones clínicas en el 2026

Desafortunadamente en Colombia el INVIMA no se ha pronunciado al respecto al momento de la escritura de este artículo sobre su posición regulatoria frente a los sistemas de soporte a la decisión clínica pero en Estados Unidos la FDA tiene un

marco regulatorio establecido y para que un software no sea regulado, es decir para que no sea considerado un software como dispositivo médico (SaMD) debe estar destinado a profesionales de la salud (HCP) y cumplir con los cuatro criterios siguientes simultáneamente (1). Las **entradas de datos *(inputs)*** no deben adquirir, procesar ni analizar imágenes médicas (como TAC, RM), datos ómicos, ni patrones/señales (como ondas de ECG o monitoreo continuo de glucosa) (2). **Los tipos de información** usados en los sistemas de soporte a la decisión clínica deben haber sido procesados previamente por un ser humano (como notas de historia clínica, informes de resultados de ayudas diagnosticas, datos de laboratorio, demografía entre otros) o bibliografía médica (guías clínicas, estudios revisados por pares) (3). **El propósito** debe ser apoyar o dar recomendaciones al profesional, no reemplazar su juicio. No debe proporcionar una directiva específica de diagnóstico o tratamiento que el médico deba seguir ciegamente (4). Deben ser **transparentes** y permitir al profesional revisar independientemente los fundamentos y la fuente de la recomendación. El médico no debe tener que confiar en la recomendación sin entender de dónde viene (evitar el efecto "caja negra") además, aunque proporcione explicación la situación clínica debe permitir analizar el output, y excluir situaciones de emergencia o criticas donde no se puede analizar la base de la recomendación de la IA. Para comprender mejor el concepto proporcionó algunos ejemplos clave y su relación la guía actual de la **IMDRF** (3) (ver **Tabla 4**).

Non-Device CDS (No Regulado):

- Software que sugiere interacciones droga-droga o contraindicaciones de alergias.
- Herramientas que unen datos del paciente con guías clínicas para sugerir una lista de tratamientos posibles para que el médico elija.
- Recordatorios de cuidados preventivos (ej. vacunas) basados en el historial médico.

Device CDS (Regulado):

- Algoritmos que predicen sepsis o shock en tiempo real para alertar al médico (tiempo crítico).
- Análisis de lunares mediante fotos para determinar riesgo de melanoma (análisis de imagen).
- Software que planifica dosis de insulina para que el paciente las use directamente (uso por paciente).
- Sistemas que analizan ondas de ECG para diagnosticar arritmias (análisis de señales/patrones).

Tabla 4. Categorización de riesgo (IMDRF) y su relación con la guía FDA en sistemas de soporte a la decisión clínica

Estado de salud del paciente (situación o condición)	Tratar o diagnosticar (*treat or diagnose*)	Guiar el manejo clínico (*drive clinical management*)	Informar el manejo clínico (*inform clinical management*)
Crítico	IV *(Dispositivo)*	III *(Dispositivo)*	II *(Generalmente Dispositivo)*
Serio	III *(Dispositivo)*	II *(Dispositivo)*	I *(Potencial No Dispositivo)*
No Serio	II *(Dispositivo)*	I (Dispositivo)	I *(No Dispositivo)*

Conclusiones y desenlace del caso clínico

Análisis de la evidencia con la guía de lectura crítica

Para responder a la pregunta inicial: ¿Considera usted que sería útil para su clínica implementar un programa de tamización de cáncer de mama apoyado en este software con Inteligencia Artificial? Aplicamos las fases de validación y las guías de lectura crítica presentadas en el manual:

1. Fase de desarrollo de la IA

El estudio MASAI es un Ensayo Controlado Aleatorizado (RCT), por lo que se ubica en la fase de **Seguridad y Efectividad a Gran Escala** (Fase 4 de DECIDE-AI). Este es el nivel de evidencia más alto que se puede esperar para una intervención médica.

Pregunta Clave: ¿El ensayo cumple con los estándares del "Gold Standard" (CONSORT-AI)?

Análisis: Sí. El MASAI es un RCT en donde el resultado era el diagnóstico final por patología. Demuestra un impacto en un resultado clínico sustituto importante (detección de cáncer y reducción de cánceres de intervalo) y en la eficiencia operativa (reducción de la carga laboral).

Limitación de transparencia: La IA, *Transpara*, es una "caja negra" comercial (Ítem de Arquitectura/Software en **Tabla 3**), lo que significa que, aunque los resultados sean impresionantes, el mecanismo de decisión y sus posibles sesgos internos no pueden ser auditados de forma independiente dado que no comparten el código fuente.

2. Análisis de sesgos

Sesgo de selección: El estudio se realizó en una población geográfica y racialmente homogénea (Suecia). **La pobreza de datos** (sesgo algorítmico) es una preocupación, ya que el rendimiento demostrado en este grupo no garantiza el mismo desempeño en poblaciones latinas, afroamericanas o con diferentes densidades mamarias, lo que limita la **generalización** del modelo (aplicación a mi contexto). La **validación externa** en el contexto colombiano es esencial antes de la implementación.

Sesgo de Automatización: Dado que el grupo de intervención era **IA + Humano**, existe un riesgo latente de que los radiólogos hayan confiado excesivamente en el puntaje de alto riesgo de la IA, lo que podría haber contribuido a la tasa de detección superior, pero también plantea la pregunta de qué tan reproducible es si el usuario humano es diferente (Elemento 7 de DECIDE-AI: Factores Humanos).

Sí, la evidencia del estudio MASAI es robusta para **considerar seriamente** la implementación del software de IA que ha demostrado no inferioridad dado que a IA logró una mayor detección sin comprometer la seguridad (tasa de falsos positivos idéntica y VPP mejorado).

La reducción significativa en la carga de trabajo del radiólogo es un factor decisivo para el sistema de salud en términos de costo-efectividad y optimización de recursos. Sin embargo debemos darle una mano con herramientas de la medicina traslacional porque a pesar de la evidencia es buena, el internista debe tener cautela y buscar **validación en población local (PROBAST+AI):** Es obligatorio realizar un estudio de **validación externa local** o una implementación piloto monitoreada para asegurar que el rendimiento y la ausencia de **sesgo algorítmico** (equidad) se mantengan en la población además de siempre tener en cuenta la **transparencia regulatoria al** cumplir el criterio de análisis de imágenes clasifica como un dispositivo (Riesgo IMDRF Clase III) por lo que se debe verificar la aprobación como *software as medical device* (SaMD) por parte de las entidades regulatorias (FDA y, para nosotros aplica por supuesto, INVIMA). Otra crítica no menor es que el estudio MASAI es que se enfoca en desenlaces intermedios como la eficiencia del radiólogo y la detección de cáncer, **no ofrece datos primarios sobre la reducción de mortalidad por cáncer de mama o los efectos a largo plazo de la IA en la calidad de vida de las pacientes**, que son los desenlaces más relevantes para el internista , por otro lado para el Sistema de Salud es llamativa el aumento de efectividad en un contexto de escasez de recursos pero se requieren, por supuesto, estudios de costo-efectividad antes de su implementación generalizada además de la planeación de la traslación tecnológica con marcos como el expuesto en el manual (RE-AIM).

Bibliografía

1. Lång K, Josefsson V, Larsson A-M, Larsson S, Högberg C, Sartor H, et al. 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 [Internet]. 2023 Aug;24(8):936–44. Available from: http://dx.doi.org/10.1016/S1470-2045(23)00298-X

2. Gommers J, Hernström V, Josefsson V, Sartor H, Schmidt D, Hjelmgren A, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet [Internet]. 2026 Jan 31;407(10527):505–14. Available from: http://dx.doi.org/10.1016/S0140-6736(25)02464-X
3. Website [Internet]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device
4. Espitia TE, Arcila EM, Junca AP. Texto de Medicina Interna. Aprendizaje basado en problemas. Segunda edición. Tomo I y II [Internet]. Distribuna Editorial Médica; 2023. 3805 p. Available from: https://play.google.com/store/books/details?id=kJUkEQAAQBAJ
5. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med [Internet]. 2022 May;28(5):924–33. Available from: http://dx.doi.org/10.1038/s41591-022-01772-9
6. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ [Internet]. 2025 Mar 24;388:e082505. Available from: http://dx.doi.org/10.1136/bmj-2024-082505
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ [Internet]. 2024 Apr 16;385:e078378. Available from: http://dx.doi.org/10.1136/bmj-2023-078378
8. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med [Internet]. 2025 Oct;31(10):3283–9. Available from: http://dx.doi.org/10.1038/s41591-025-03953-8
9. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med [Internet]. 2020 Sept;26(9):1351–63. Available from: http://dx.doi.org/10.1038/s41591-020-1037-7
10. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health [Internet]. 2020 Oct;2(10):e537–48. Available from: http://dx.doi.org/10.1016/S2589-7500(20)30218-1
11. Glasgow RE, Harden SM, Gaglio B, Rabin BA, Ory MG, Smith ML, et al. Use of the RE-AIM Framework: Translating Research to Practice with Novel Applications and Emerging Directions [Internet]. Frontiers Media SA; 2021. 198 p. Available from: https://books.google.com/books/about/Use_of_the_RE_AIM_Framework_Translating.html?hl=&id=hxdDEAAAQBAJ
12. Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation [Internet]. World Health Organization; 2021. 104 p. Available from: https://play.google.com/store/books/details?id=jMJqEAAAQBAJ
13. Asai A, He J, Shao R, Shi W, Singh A, Chang JC, et al. Synthesizing scientific literature with retrieval-augmented language models. Nature [Internet]. 2026 Feb;650(8103):857–63. Available from: http://dx.doi.org/10.1038/s41586-025-10072-4
14. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med [Internet]. 2015 June 2;162(11):777–84. Available from: http://dx.doi.org/10.7326/M14-2385
15. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med [Internet]. 2023 Jan;29(1):14–5. Available from: http://dx.doi.org/10.1038/s41591-022-02139-w
16. Holst D, Moenck K, Koch J, Schmedemann O, Schüppstuhl T. Transparent Reporting of AI in systematic literature reviews: Development of the PRISMA-trAIce checklist. JMIR AI [Internet]. 2025 Dec 10;4(v4i1e80247):e80247. Available from: http://dx.doi.org/10.2196/80247

Tabla 3. Comparación de Guías de lectura crítica de la literatura de estudios que usan métodos de aprendizaje de maquina en Medicina

Categoría de Ítem	TRIPOD+AI (Pronóstico)	STARD-AI (Diagnóstico)	CONSORT-AI (Ensayos Clínicos)	DECIDE-AI (Eval. Temprana)	PROBAST+AI (Riesgo de Sesgo)	JAMA ML Guide (Lectura Crítica)
Identificación (Título)	Ítem 1: Identificar como desarrollo/ validación de modelo usando IA/ML.	Ítem 1: Identificar como estudio de precisión diagnóstica usando IA.	Ítem 1a: Indicar que la intervención involucra IA.	Elemento 1: Identificar como evaluación clínica temprana de IA.		
Resumen (Abstract)	Ítem 2: Resumen estructurado (objetivos, datos, modelo, desempeño).	Ítem 2: Resumen que incluya el sistema de IA y desempeño.	Ítem 1b: Resumen que especifique la intervención de IA.	Elemento I: Resumen (algoritmo, ámbito, seguridad).		
Uso Previsto y Objetivos	Ítem 4: Objetivos y rol del modelo en el cuidado de salud.	Ítem 4: Objetivos del estudio y rol del sistema de IA.		Elemento 2: Descripción de la condición, usuarios y población.		
Fuentes de Datos	Ítem 5: Fuentes de datos y fechas de reclutamiento.	Ítem 6: Fuentes de datos para sets de entrenamiento y prueba.		Elemento 5: Diversidad y representatividad de los datos.	Dominio 1: ¿Fueron apropiadas las fuentes de datos?	¿Se usaron datos independientes para validación?
Arquitectura / Software	Ítem 11: Descripción del algoritmo, arquitectura y preprocesamiento.	Ítem 12: Versión de IA, arquitectura y software utilizado.	Ítem 5: Versión de la IA, hardware e instrucciones de uso.	Elemento Nuevo 1: Arquitectura, funcionalidad e interpretabilidad.	Dominio 4: ¿Fue adecuado el desarrollo y ajuste del modelo?	¿Es apropiado el tipo de modelo para los datos?

Continúa en la siguiente página.

Tabla 3. Comparación de Guías de lectura crítica de la literatura de estudios que usan métodos de aprendizaje de maquina en Medicina (Continuación)

Entrenamiento / Ajuste	Ítem 10b: Manejo de hiperparámetros y optimización.	Ítem 12b: Detalles del entrenamiento del algoritmo.				¿Cómo se ajustaron los hiperparámetros?
Métricas de Desempeño	Ítem 16: Medidas de desempeño (calibración y discriminación).	Ítem 20: Métricas de precisión diagnóstica (sensibilidad, etc.).	Ítem 19: Reportar el desempeño y errores de la IA.	Elemento 14: Resultados de desempeño y métricas de seguridad.	Dominio 4: Evaluación de calibración y discriminación.	¿Qué tan precisos son los resultados?
Análisis de Error / Seguridad	*Ítem 22: Comparación con otros modelos o estándares.*		*Ítem 19: Descripción de fallos y casos de daño (safety).*	*Elemento 15: Discusión de errores, fallos y mitigación.*		
Ética y Transparencia	*Ítem 26/27: Disponibilidad de datos y código.*	Ítem 27-30: Financiamiento, registro y protocolo.		*Elemento Nuevo 2: Estándares éticos y principios de bioética.*		
Interpretación Clínica				*Elemento 7: Evaluación de factores humanos y usuarios.*		*¿Cómo se aplican los resultados al paciente?*