Ciencia y Educación
(L-ISSN: 2790-8402 E-ISSN: 2707-3378)
Vol. 7 No. 11.1
Edición Especial XI 2026

EFECTOS DEL USO DE LA INTELIGENCIA ARTIFICIAL SOBRE EL RAZONAMIENTO CLÍNICO EN ESTUDIANTES DE MEDICINA: REVISIÓN SISTEMÁTICA
EFFECTS OF THE USE OF ARTIFICIAL INTELLIGENCE ON CLINICAL REASONING IN MEDICAL STUDENTS: A SYSTEMATIC REVIEW

Autores: [1]Lissette Gardenia Hoyos Sanmartín y [2]Eugenio Alejandro Ortega García.
[1]ORCID ID: https://orcid.org/0009-0002-7534-795X
[2]ORCID ID: https://orcid.org/0000-0003-2537-3965
[1]E-mail de contacto: lhoyoss2@unemi.edu.ec
[2]E-mail de contacto: eortegag2@unemi.edu.ec
Afiliación: [1*2*]Universidad Estatal de Milagro, (Ecuador).
Artículo recibido: 18 de Octubre del 2026.
Artículo revisado: 25 de Octubre del 2026.
Artículo aprobado: 25 de Octubre del 2026.
[1]Médico general, graduada de la Universidad Técnica de Machala, (Ecuador). Maestrante de la Maestría en Educación con mención en Docencia e Investigación en Educación Superior de la Universidad Estatal de Milagro, (Ecuador).
[2]Licenciado en Ciencias de la Educación mención Educación Básica, graduado de la Universidad Estatal de Milagro, (Ecuador). Magíster en Gerencia Educativa de la Universidad Estatal de Milagro, (Ecuador).

Resumen

El objetivo de la revisión sistemática fue analizar los efectos del uso de herramientas de inteligencia artificial sobre el razonamiento clínico en estudiantes de medicina, mediante la integración de estudios primarios comparativos. El estudio se desarrolló conforme a la declaración *Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020*. La búsqueda se realizó en bases de información científica de amplia cobertura y reconocimiento académico, complementada con la revisión de sitios editoriales. Se incluyeron estudios comparativos con estudiantes de medicina, intervenciones de educación médica basadas en inteligencia artificial, comparadores identificables y desenlaces relacionados con el razonamiento clínico. El riesgo de sesgo se evaluó mediante instrumentos específicos según el diseño del estudio y los hallazgos se integraron mediante síntesis narrativa siguiendo las orientaciones de *Synthesis Without Meta-analysis.* Se incluyeron 30 estudios. Diecinueve mostraron efectos predominantemente favorables, siete resultados comparables, tres efectos mixtos o condicionados y uno presentó un efecto desfavorable. Los beneficios se concentraron en actividades estructuradas con práctica activa y retroalimentación, sin demostrar una superioridad uniforme frente a los métodos convencionales. La exposición a contenido clínicamente inexacto generado por estas herramientas redujo la precisión diagnóstica y alteró la calibración de la confianza. El predominio de estudios con riesgo de sesgo moderado o alto limitó la interpretación causal de los hallazgos. Estas tecnologías pueden fortalecer componentes del razonamiento clínico cuando complementan propuestas pedagógicas estructuradas y supervisadas. No sustituyen la enseñanza docente, la práctica clínica supervisada ni el criterio profesional; su incorporación requiere verificar la información, reconocer errores y justificar las decisiones clínicas.

Palabras clave: Inteligencia artificial, Razonamiento clínico, Estudiantes de medicina, Educación médica, Aprendizaje clínico.

Abstract

The objective of the systematic review was to analyze the effects of the use of artificial intelligence on clinical reasoning among medical students through the integration of comparative primary studies. The study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement. The search was conducted in widely recognized scientific databases with broad coverage, supplemented by searches of publisher websites. Comparative studies involving medical

students, artificial intelligence-based medical education interventions, identifiable comparators, and outcomes related to clinical reasoning were included. Risk of bias was assessed using specific instruments according to study design, and the findings were integrated through narrative synthesis following the Synthesis Without Meta-analysis guidelines. Thirty studies were included. Nineteen showed predominantly favorable effects, seven reported comparable results, three showed mixed or conditional effects, and one reported an unfavorable effect. Benefits were concentrated in structured activities involving active practice and feedback, without demonstrating uniform superiority over conventional methods. Exposure to clinically inaccurate content generated by these tools reduced diagnostic accuracy and altered confidence calibration. The predominance of studies with moderate or high risk of bias limited the causal interpretation of the findings. These technologies may strengthen components of clinical reasoning when they complement structured and supervised educational approaches. They do not replace faculty-led teaching, supervised clinical practice, or professional judgment; their implementation requires verification of information, recognition of errors, and justification of clinical decisions.

Keywords: Artificial intelligence, clinical reasoning, medical students, medical education, systematic review.

Sumário

O objetivo da revisão sistemática foi analisar os efeitos do uso de ferramentas de inteligência artificial sobre o raciocínio clínico em estudantes de medicina, por meio da integração de estudos primários comparativos. O estudo foi conduzido de acordo com a declaração *Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020.* A busca foi realizada em bases de dados científicas de ampla cobertura e reconhecimento acadêmico, complementada pela consulta a sites das editoras. Foram incluídos estudos comparativos com estudantes de medicina, intervenções de educação médica baseadas em inteligência artificial, comparadores identificáveis e desfechos relacionados ao raciocínio clínico. O risco de viés foi avaliado por meio de instrumentos específicos conforme o delineamento de cada estudo, e os achados foram integrados mediante síntese narrativa, seguindo as orientações de *Synthesis Without Meta-analysis.* Foram incluídos 30 estudos. Dezenove apresentaram efeitos predominantemente favoráveis, sete resultados comparáveis, três efeitos mistos ou condicionados e um apresentou efeito desfavorável. Os benefícios concentraram-se em atividades estruturadas com prática ativa e *feedback*, sem demonstrar superioridade uniforme em relação aos métodos convencionais. A exposição a conteúdo clinicamente impreciso gerado por essas ferramentas reduziu a precisão diagnóstica e alterou a calibração da confiança. O predomínio de estudos com risco de viés moderado ou alto limitou a interpretação causal dos achados. Essas tecnologias podem fortalecer componentes do raciocínio clínico quando complementam propostas pedagógicas estruturadas e supervisionadas. Não substituem o ensino ministrado por docentes, a prática clínica supervisionada nem o julgamento profissional; sua incorporação requer a verificação das informações, o reconhecimento de erros e a justificativa das decisões clínicas.

Palavras-chave: Inteligência artificial, raciocínio clínico, estudantes de medicina, educação médica, revisão sistemática.

Introducción

La inteligencia artificial (IA) comprende sistemas computacionales que procesan información, reconocen patrones y generan respuestas orientadas a tareas específicas. Su incorporación en la educación médica amplía las oportunidades de práctica, puede facilitar la retroalimentación oportuna y ofrece apoyo durante el aprendizaje clínico. En estudiantes de medicina, estas

posibilidades adquieren especial relevancia, dado que el uso de la IA se ha asociado con mejoras en la comunicación, la atención al paciente, los conocimientos y otras habilidades vinculadas con el desarrollo de competencias profesionales (Shen et al., 2026).

El razonamiento clínico constituye uno de los núcleos de la formación médica, porque permite transformar los datos obtenidos del paciente en hipótesis diagnósticas, decisiones terapéuticas y planes de manejo fundamentados. No se limita a identificar síntomas, signos o recordar contenidos; exige comprender y procesar la información proporcionada por el paciente, reconocer patrones clínicos y evaluar la evidencia para establecer un diagnóstico y determinar un plan de tratamiento apropiado (Brügge et al., 2024).

En los estudios primarios seleccionados, la IA no aparece como una intervención única. Sus formas de incorporación comprenden pacientes estandarizados virtuales (Shen et al., 2026), pacientes simulados mediante IA basados en modelos de lenguaje de gran escala (LLM, por sus siglas en inglés) (Brügge et al., 2024), pacientes simulados desarrollados con un transformador generativo preentrenado (GPT, por sus siglas en inglés), utilizado a través de ChatGPT, plataforma conversacional que permite interactuar con este tipo de modelo (Wang et al., 2025), plataformas multimodales basadas en IA (J. Xu et al., 2026), estrategias de aprendizaje basado en problemas (ABP) asistidas por IA (Gu et al., 2026), entre otras posibilidades educativas que seguirán diversificándose con el desarrollo tecnológico. En conjunto, estas modalidades ofrecen apoyo en distintas etapas del aprendizaje clínico, siempre como complemento de la enseñanza docente y del criterio profesional. Entre estas aplicaciones, los pacientes simulados mediante IA ofrecen entornos de práctica para desarrollar habilidades clínicas antes del contacto con pacientes reales. Su uso facilita la repetición de entrevistas, la retroalimentación inmediata, el fortalecimiento de la recopilación de información, el razonamiento clínico y la comunicación (Wang et al., 2025).

Los efectos descritos varían según la intervención y la forma de evaluación, con resultados que comprenden desde desempeños comparables con métodos convencionales (Dang et al., 2026) hasta mejoras en distintos componentes de las habilidades clínicas evaluadas después de prácticas adicionales con estas herramientas (McCarrick et al., 2025). Los resultados favorables fueron más evidentes cuando la IA se integró en procesos dirigidos al desarrollo del razonamiento y las habilidades clínicas. La combinación de LLM con métodos tradicionales aumentó la precisión diagnóstica (Roemer et al., 2026). Mientras la instrucción asistida por IA mantuvo resultados semejantes en conocimientos teóricos y favoreció las habilidades clínicas (Ba et al., 2024). La utilidad de los sistemas de apoyo a la decisión clínica también estuvo condicionada por la calidad de la información ingresada, con mayor exactitud cuando los hallazgos se expresaron mediante términos médicos precisos (Yanagita et al., 2025).

Algunas comparaciones incluidas en esta revisión evidenciaron que la IA no siempre superó a las formas convencionales de enseñanza y evaluación clínica. Al comparar la retroalimentación generada por IA con la escrita por expertos, ambas produjeron resultados generales semejantes en las habilidades de

razonamiento clínico; sin embargo, la retroalimentación escrita por expertos fue más eficaz para resolver casos de mayor complejidad (Çiçek et al., 2025). Durante evaluaciones clínicas con pacientes reales, el uso de IA generativa tampoco produjo mejoras significativas en el desempeño clínico y el razonamiento diagnóstico frente a la evaluación clínica convencional (Saloojee et al., 2026). En conjunto, estos hallazgos indican que el valor de la IA depende de la naturaleza de la actividad, su complejidad y el contexto de aplicación. Su disponibilidad, por sí sola, no garantiza un razonamiento clínico más profundo o preciso (Çiçek et al., 2025; Saloojee et al., 2026).

Otros estudios examinaron la contribución de la simulación asistida por IA al desarrollo de habilidades clínicas prácticas. El entrenamiento con pacientes simulados mediante IA favoreció el desarrollo de habilidades clínicas y mejoró el desempeño estudiantil (Gao et al., 2026). Al comparar pacientes estandarizados virtuales impulsados por IA con pacientes estandarizados convencionales, ambas modalidades mejoraron el desempeño clínico y produjeron resultados semejantes, sin una ventaja significativa de la IA (Dang et al., 2026). Estos hallazgos indican que la simulación constituye un recurso útil para el entrenamiento clínico, aunque la mejoría observada no siempre puede atribuirse exclusivamente al componente tecnológico (Dang et al., 2026; Gao et al., 2026).

El uso de IA en educación médica exige considerar sus beneficios y posibles riesgos. Las explicaciones clínicas generadas por IA produjeron efectos diferentes según su exactitud: la información incorrecta redujo significativamente la precisión diagnóstica, mientras que la información correcta no produjo mejoras significativas frente a la evaluación clínica convencional (Teng et al., 2026). En la resolución de preguntas médicas, las recomendaciones generadas por IA mejoraron modestamente la precisión de las respuestas, aunque las recomendaciones erróneas también influyeron en la toma de decisiones e indujeron algunos errores (Kirpalani y Wang, 2026). En la interpretación de estudios de imagen, el entrenamiento asistido por IA mejoró el desempeño diagnóstico y parte del beneficio se mantuvo después de retirar la herramienta (Tee et al., 2026). Estos hallazgos muestran que la IA puede producir efectos favorables, neutrales o desfavorables según la exactitud del contenido y la valoración crítica del estudiante (Kirpalani y Wang, 2026; Tee et al., 2026; Teng et al., 2026).

Las diferencias metodológicas impiden asumir un efecto uniforme entre los estudios. La retroalimentación individual mediante IA no superó a la explicación experta general, aunque su utilización se relacionó con mejores resultados (Rother et al., 2026). En una cohorte histórica, la IA mejoró la enseñanza entre pares sin aislar su efecto tecnológico (Yu et al., 2026). La diversidad de intervenciones, comparadores y desenlaces exige una síntesis estructurada e interpretación diferenciada de la evidencia (Rother et al., 2026; Yu et al., 2026).

Para examinar esta evidencia, se realizó una revisión sistemática con síntesis narrativa estructurada, sin metaanálisis, de estudios primarios con intervenciones educativas comparativas. Los resultados fueron favorables, comparables, mixtos, condicionados o desfavorables, según las diferencias metodológicas y educativas entre los estudios.

Ante esta evidencia, surge la siguiente pregunta de investigación: ¿qué efectos tiene el uso de herramientas de IA sobre el razonamiento clínico en estudiantes de medicina? En consecuencia, el objetivo de esta revisión sistemática fue analizar los efectos del uso de herramientas de IA sobre el razonamiento clínico en estudiantes de medicina, mediante la integración de estudios primarios comparativos publicados entre enero de 2022 y julio de 2026.

Materiales y Métodos

Se realizó una revisión sistemática con enfoque de síntesis narrativa estructurada, sin metaanálisis, orientada a analizar los efectos del uso de herramientas de IA sobre el razonamiento clínico en estudiantes de medicina. La revisión integró estudios primarios con intervenciones educativas comparativas. Los resultados fueron examinados mediante comparación, agrupación e interpretación crítica.

El proceso de identificación, selección y presentación de la evidencia se estructuró de acuerdo con la declaración *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA 2020), empleada para documentar de manera transparente las etapas de la revisión (Page et al., 2021). La síntesis siguió las orientaciones de *Synthesis Without Meta-analysis* (SWiM) para agrupar, comparar e integrar narrativamente los resultados sin metaanálisis (Campbell et al., 2020). La cantidad de investigaciones que compartieron el mismo diseño, comparador, instrumento y componente del razonamiento clínico evaluado fue limitada para establecer agrupaciones estadísticas consistentes. La síntesis narrativa permitió conservar el significado clínico y educativo de cada hallazgo, evitando presentar una estimación conjunta que pudiera ocultar diferencias relevantes entre las intervenciones evaluadas.

La revisión respondió a la pregunta: ¿qué efectos tiene el uso de herramientas de IA sobre el razonamiento clínico en estudiantes de medicina? Su delimitación se organizó mediante los componentes población, intervención, comparador y resultado (PICO). La población estuvo constituida por estudiantes de medicina de pregrado, incluidos internos que permanecían vinculados con su formación universitaria.

La intervención comprendió el uso directo de IA por los estudiantes o la exposición a recursos educativos generados mediante estas herramientas. Los comparadores admitidos fueron métodos tradicionales de enseñanza, retroalimentación escrita por expertos, pacientes estandarizados convencionales, ausencia de asistencia tecnológica, evaluación clínica convencional o desempeño previo de los mismos participantes. El razonamiento clínico fue entendido como un proceso multidimensional que integra la recopilación y el análisis de información, la formulación de hipótesis, el diagnóstico diferencial, el criterio clínico, la toma de decisiones y la planificación terapéutica. También se consideraron las actividades dirigidas a obtener y organizar información clínica cuando contribuían a sustentar decisiones diagnósticas o terapéuticas.

Se aceptaron estudios primarios comparativos, incluidos ensayos aleatorizados con diseño paralelo, por conglomerados o cruzado; estudios cuasiexperimentales y comparativos no aleatorizados; cohortes históricas; y diseños antes-después con comparador concurrente o intrapersonal. Cada investigación debía evaluar al

menos un componente del razonamiento clínico mediante una medición observable del desempeño estudiantil. Se consideraron estudios primarios publicados en revistas científicas, en inglés o español, durante el periodo comprendido entre enero de 2022 y el 31 de julio de 2026. Este intervalo permitió reunir evidencia reciente en un campo caracterizado por su rápida evolución tecnológica. Los artículos debían disponer de texto completo, incluir estudiantes de medicina, aplicar una intervención basada en IA y presentar un comparador identificable.

Se excluyeron revisiones, metaanálisis, editoriales, cartas, comentarios, investigaciones realizadas con médicos titulados, residentes o especialistas; estudios sin comparador; intervenciones sin correspondencia con el propósito educativo de la revisión; y trabajos que no evaluaron algún componente del razonamiento clínico mediante resultados observables. Por tratarse de una revisión sistemática, no se realizó un muestreo de participantes; la muestra documental quedó determinada por los estudios que cumplieron los criterios de elegibilidad establecidos.

La búsqueda bibliográfica se realizó en bases de información científica de amplia cobertura y reconocimiento académico, complementada con la revisión de sitios editoriales, y se desarrolló entre febrero y julio de 2026. Para cerrar y verificar el registro, las estrategias definitivas de PubMed y Scopus se ejecutaron y exportaron el 17 de agosto de 2026, manteniendo como límite de publicación el periodo comprendido entre el 1 de enero de 2022 y el 31 de julio de 2026. De manera complementaria, se revisó la información pública sobre la indexación de las revistas que publicaron los estudios incluidos en ERIC, Web of Science Core Collection, Embase, PubMed y Scopus, cuando correspondía. Se observó que algunas revistas estaban incluidas en más de una fuente bibliográfica. Esta comprobación tuvo carácter descriptivo y no formó parte de las estrategias definitivas de búsqueda ni del proceso de identificación de registros. Las bases de datos consultadas mediante estrategias estructuradas fueron PubMed y Scopus.

La búsqueda se organizó mediante tres grupos conceptuales relacionados con la intervención, la población y el desenlace. El primer grupo comprendió IA, IA generativa, ChatGPT, modelos de lenguaje, sistemas de apoyo a la decisión clínica y pacientes virtuales. El segundo incluyó estudiantes de medicina, educación médica de pregrado e internos. El tercero reunió razonamiento clínico, razonamiento diagnóstico, toma de decisiones clínicas, precisión diagnóstica y anamnesis. Los términos equivalentes se combinaron con OR y los grupos con AND.

En PubMed se ejecutó la siguiente estrategia: ((“artificial intelligence”[Title/Abstract] OR “generative artificial intelligence”[Title/Abstract] OR ChatGPT[Title/Abstract] OR “large language model”[Title/Abstract] OR “clinical decision support system”[Title/Abstract] OR “virtual patient”[Title/Abstract]) AND (“medical students”[Title/Abstract] OR “undergraduate medical education”[Title/Abstract] OR “medical interns”[Title/Abstract]) AND (“clinical reasoning”[Title/Abstract] OR “diagnostic reasoning”[Title/Abstract] OR “clinical decision-making”[Title/Abstract] OR “diagnostic accuracy”[Title/Abstract] OR “history-taking”[Title/Abstract])) AND (“2022/01/01”[Date - Publication] :

"2026/07/31"[Date - Publication]). La consulta identificó 206 registros. En Scopus se utilizó la expresión TITLE-ABS-KEY(("artificial intelligence" OR "generative artificial intelligence" OR ChatGPT OR "large language model" OR "clinical decision support system" OR "virtual patient") AND ("medical students" OR "undergraduate medical education" OR "medical interns") AND ("clinical reasoning" OR "diagnostic reasoning" OR "clinical decision-making" OR "diagnostic accuracy" OR "history-taking")) AND PUBYEAR > 2021 AND PUBYEAR < 2027. Antes de la exportación se mantuvo el límite de publicación hasta el 31 de julio de 2026. El archivo final exportado en formato de valores separados por comas (CSV) contuvo 372 registros.

Las búsquedas exploratorias anteriores y la comprobación en sitios editoriales permitieron recuperar nueve informes incluidos que no aparecieron en las ecuaciones definitivas de PubMed y Scopus. Los títulos, autores, años e identificadores de objeto digital (DOI) fueron verificados en los sitios editoriales de BMC y Nature Portfolio, pertenecientes a Springer Nature, así como en JMIR Publications, Taylor & Francis, Dove Medical Press, ScienceDirect y Frontiers.

La clasificación editorial de las revistas se verificó en *SCImago Journal & Country Rank* tras determinar la elegibilidad de los estudios, como descriptor complementario de su visibilidad y reconocimiento internacional. El cuartil no se utilizó como criterio de inclusión o exclusión ni como medida de calidad metodológica, evaluada mediante herramientas específicas de riesgo de sesgo. Los resultados de PubMed y Scopus se reunieron en una matriz común a partir del listado de PubMed y del archivo CSV exportado desde Scopus. La deduplicación se efectuó mediante DOI, identificador de PubMed (PMID), título normalizado, autores y año de publicación. Se identificaron 153 coincidencias por DOI y una coincidencia adicional por PMID y título, para un total de 154 duplicados. No se utilizaron herramientas automatizadas para decidir la elegibilidad; la revisión de títulos, resúmenes y textos completos fue realizada por los autores según los criterios previamente definidos.

La investigadora principal examinó el cumplimiento de los criterios de población, intervención, comparador, desenlace, diseño, idioma y periodo de publicación. Un segundo revisor comprobó las decisiones de inclusión, las razones de exclusión y la correspondencia de los estudios con la pregunta planteada. Las diferencias fueron resueltas mediante consenso y nueva lectura del texto completo. Las razones de exclusión durante la lectura a texto completo se agruparon según el principal criterio de elegibilidad incumplido. Los resultados del proceso de selección se presentan en la sección "Selección de estudios" y en la Figura 1.

La información se recopiló mediante una matriz diseñada específicamente para esta revisión, la cual incluye los campos detallados en la Tabla 1. Se registraron autor, año, país, diseño, muestra, población, intervención con IA, comparador, propósito del estudio y resultado principal. La extracción fue realizada por la investigadora principal, quien contrastó cada elemento con el texto completo antes de incorporarlo a las tablas de resultados. Los resultados principales se conservaron en la medida comunicada por cada estudio, incluidas medias y desviaciones

estándar, proporciones, diferencias entre grupos, *odds ratios*, tamaños del efecto, intervalos de confianza y valores de significación cuando estuvieron disponibles. No se estableció una medida común porque los estudios utilizaron desenlaces e instrumentos heterogéneos. Cada resultado se examinó según el componente del razonamiento clínico evaluado, su dirección frente al comparador y el riesgo de sesgo del estudio. No se recalcularon efectos ni se transformaron escalas para producir comparaciones estadísticas artificiales.

Cuando un estudio comunicó varios resultados, para la síntesis se seleccionó aquel relacionado de forma más directa con el razonamiento clínico. En las puntuaciones globales que integraban varias habilidades se priorizó el dominio con mayor correspondencia con la interpretación clínica, el diagnóstico o la toma de decisiones. Cuando se evidenciaban evaluaciones inmediatas y diferidas, se priorizó la medición inmediata posterior a la intervención para la comparación principal. La evaluación diferida se conservó como información complementaria sobre la retención del aprendizaje. Este procedimiento evitó contabilizar varias veces a los mismos participantes dentro de una misma interpretación.

Los datos no comunicados se registraron como no informados y no se efectuaron imputaciones. No se contactó a los autores para solicitar información adicional ni se estimaron efectos estadísticos conjuntos. La información extraída se organizó en tres tablas. La Tabla 1 presenta las características PICO y metodológicas de los estudios incluidos; la Tabla 2 agrupa los resultados según la dirección y el alcance del efecto observado; y la Tabla 3 presenta la valoración del riesgo de sesgo. Esta organización permite relacionar las características de las intervenciones con sus resultados y con la confianza metodológica concedida a cada hallazgo.

El riesgo de sesgo fue evaluado inicialmente por la investigadora principal y revisado posteriormente por un segundo autor. La valoración se efectuó sobre el desenlace relacionado de forma más directa con el razonamiento clínico. Las diferencias entre los revisores fueron resueltas mediante consenso y nueva lectura del artículo. Los ensayos aleatorizados fueron examinados mediante *Cochrane Risk of Bias 2* (RoB 2), incluidas las versiones correspondientes a diseños cruzados o aleatorizados por conglomerados (Sterne et al., 2019). Los estudios comparativos no aleatorizados se valoraron mediante *Risk Of Bias In Non-randomized Studies of Interventions* (ROBINS-I) (Sterne et al., 2016). Los diseños cuasiexperimentales antes y después se revisaron mediante el instrumento del Joanna Briggs Institute (JBI) para estudios cuasiexperimentales (Barker et al., 2024).

Para facilitar la comparación entre instrumentos, las valoraciones se presentaron en una escala común de riesgo bajo, moderado o alto. En RoB 2, la categoría original de algunas preocupaciones se expresó como riesgo moderado únicamente para la presentación comparativa; las categorías de riesgo bajo y alto conservaron su denominación. En ROBINS-I, las categorías grave y crítica se integraron como riesgo alto. En JBI, la valoración global se estableció a partir de los dominios metodológicos relevantes, sin aplicar una puntuación numérica. La valoración consideró la asignación de los participantes, la comparabilidad inicial, las desviaciones respecto

de la intervención, la integridad del seguimiento, la medición de los desenlaces y la correspondencia entre los análisis realizados y los resultados comunicados. Ningún estudio fue excluido exclusivamente por su nivel de riesgo; la evaluación se utilizó para graduar la confianza interpretativa concedida a sus hallazgos. Los resultados se presentan en la Tabla 3 mediante cuatro elementos: referencia, instrumento aplicado, valoración global y fundamento principal. La clasificación expresa el criterio metodológico del equipo revisor y se fundamenta en las características internas de cada investigación.

Los resultados fueron integrados mediante una síntesis narrativa estructurada conforme a SWiM. Para su presentación, los estudios se clasificaron según la dirección y el alcance del efecto observado: predominantemente favorable, comparable frente al comparador, mixto o condicionado y predominantemente desfavorable. Dentro de cada categoría se compararon las aplicaciones educativas de la IA, la población, el diseño, el comparador, el desenlace y el riesgo de sesgo. La heterogeneidad se examinó a partir de las diferencias en la población, el diseño, el comparador, el desenlace, el riesgo de sesgo, la complejidad de las actividades, la duración de las intervenciones y los instrumentos utilizados. También se consideró si la IA se aplicó de manera aislada o combinada con otras estrategias educativas que pudieran influir en el resultado.

Debido a la diversidad de escalas y mediciones, no se estableció una medida de efecto común. Los resultados se interpretaron según su dirección, alcance y consistencia, en relación con el comparador y el componente del razonamiento clínico evaluado. La interpretación también consideró el diseño y el riesgo de sesgo, sin asumir que una mayor cantidad de estudios favorables equivalía a una mayor certeza.

No se efectuaron análisis formales de sensibilidad ni de sesgo de publicación debido a la ausencia de agrupaciones homogéneas suficientes. Los estudios con riesgo alto se interpretaron con mayor cautela y no se utilizaron de manera aislada para sustentar las conclusiones. No se utilizó un sistema formal para calificar la certeza global de la evidencia. La confianza en los hallazgos se examinó de forma interpretativa mediante el diseño del estudio, el riesgo de sesgo, la consistencia entre investigaciones y la correspondencia directa de los desenlaces con la pregunta de revisión.

La revisión utilizó exclusivamente información procedente de artículos científicos publicados. No se realizaron intervenciones con seres humanos o recopilaciones de datos personales. Tampoco se accedió a información clínica identificable, muestras biológicas o conjuntos de datos restringidos. En consecuencia, no se requirió aprobación de un comité de ética o consentimiento informado. Si bien la revisión no fue registrada prospectivamente y su protocolo de trabajo no se publicó como un documento independiente, los criterios de elegibilidad, las fuentes de información, las estrategias de búsqueda, el procedimiento de selección, la extracción de datos, las herramientas para valorar el riesgo de sesgo y el método de síntesis se documentaron de forma detallada en el manuscrito. Durante la elaboración del manuscrito se utilizó ChatGPT, desarrollado por OpenAI, en las versiones disponibles en la plataforma entre febrero y agosto de 2026, como

herramienta auxiliar para organizar la información, revisar la claridad lingüística y apoyar la reformulación de borradores. La herramienta no tomó decisiones de elegibilidad, no realizó de manera autónoma la extracción de datos y no determinó las valoraciones del riesgo de sesgo. Los datos y las interpretaciones fueron contrastados con los textos completos. Las decisiones de selección, las valoraciones metodológicas, la integridad del contenido y las conclusiones corresponden a los autores. Este proceso mantuvo la correspondencia entre la evidencia examinada y su presentación en el manuscrito.

Resultados y Discusión

Las búsquedas en las bases de datos identificaron 578 registros: 206 en PubMed y 372 en Scopus. Después de eliminar 154 duplicados, 424 registros únicos fueron revisados por título y resumen. En esta etapa se excluyeron 383 registros y se solicitaron 41 informes para recuperación. Todos los textos completos estuvieron disponibles y fueron evaluados según los criterios de población, intervención, comparador, desenlace, diseño, idioma y periodo de publicación. Las búsquedas exploratorias y la comprobación en sitios editoriales aportaron nueve informes adicionales que no fueron recuperados mediante las ecuaciones definitivas de PubMed y Scopus. Los nueve fueron obtenidos y evaluados a texto completo. En conjunto, se examinaron 50 informes, de los cuales 20 fueron excluidos: cuatro por población no elegible, nueve por ausencia de comparador, cinco por corresponder a diseños no primarios y dos por evaluar intervenciones sin relación directa con el propósito de la revisión. La muestra final quedó constituida por 30 estudios primarios.

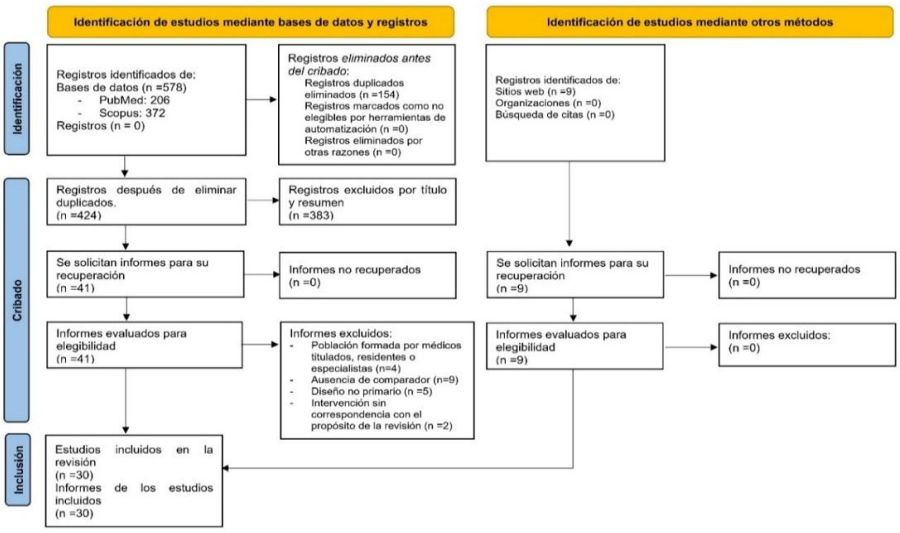


Figura 1. *Diagrama de flujo de identificación y selección de estudios conforme a PRISMA 2020.*

Fuente: Elaboración propia.

Los informes que parecían cumplir inicialmente los criterios, pero fueron excluidos después de la lectura completa, se clasificaron según un único motivo principal: población formada por médicos

titulados, residentes o especialistas; ausencia de un comparador identificable; diseño no primario; o intervención sin correspondencia con el propósito educativo de la revisión. Esta clasificación evitó la doble contabilización y mantuvo la correspondencia entre las razones de exclusión, el recuento presentado en el diagrama de flujo y la selección final. La secuencia completa del proceso de identificación, evaluación y selección de los estudios se presenta de manera esquemática en la Figura 1.

Los 30 estudios se desarrollaron en 13 países e incluyeron estudiantes de medicina desde los primeros años del pregrado hasta el internado. Predominaron los ensayos controlados aleatorizados, junto con estudios cuasiexperimentales, cohortes comparativas, diseños cruzados y estudios antes y después. Esta diversidad permitió examinar la IA en distintos momentos de la formación médica. La producción se concentró principalmente en China, con 14 estudios, seguida de Alemania con cuatro y Japón con dos. Brasil, Canadá, Irlanda, Italia, Sudáfrica, Suecia, Tailandia, Taiwán, Turquía y Vietnam estuvieron representados por una investigación cada uno.

La distribución evidencia una participación internacional amplia, aunque con mayor presencia del contexto asiático. La evidencia incluida refleja un campo de investigación en rápida expansión, con predominio de estudios recientes orientados a distintas aplicaciones de IA en la educación médica. Esta concentración confirma el creciente interés por incorporar estas herramientas en escenarios de aprendizaje clínico. Las muestras oscilaron entre 16 y 190 participantes, con tamaños muy variables entre los estudios. Esta amplitud refleja diferencias en el alcance de las investigaciones y en los contextos educativos evaluados, desde experiencias de pequeña escala hasta estudios con grupos más numerosos.

Las intervenciones incluyeron pacientes simulados, modelos conversacionales, plataformas multimodales, sistemas de apoyo a la decisión clínica, retroalimentación automatizada y recursos para interpretación de imágenes. La IA se integró en actividades de simulación, resolución de casos, práctica guiada y apoyo al aprendizaje clínico. Los comparadores estuvieron representados por métodos tradicionales de enseñanza, retroalimentación escrita por expertos, pacientes estandarizados convencionales, recursos clínicos tradicionales, evaluación clínica convencional o desempeño previo. Esta variedad permitió contrastar la IA frente a diferentes modalidades de enseñanza y apoyo clínico.

Los estudios abordaron anamnesis, razonamiento y pensamiento clínicos, precisión diagnóstica, toma de decisiones, resolución de casos, interpretación de imágenes y desempeño clínico. También variaron las formas de presentar los resultados, con puntuaciones continuas, proporciones, cambios antes y después, diferencias entre grupos y tamaños del efecto. La diversidad de los propósitos clínicos evaluados muestra que la IA fue incorporada en distintas fases del razonamiento, desde la obtención de información y la formulación diagnóstica hasta la toma de decisiones y la planificación terapéutica. La Tabla 1 reúne estas características PICO y metodológicas de manera individual para cada estudio, incluyendo autor y año, país y diseño, muestra y población, intervención con IA, comparador, propósito y resultado principal.

Tabla 1. *Características PICO y metodológicas de los estudios incluidos.*

Autor/año	País/diseño	Muestra/población	Intervención con IA / comparador	Propósito del estudio	Resultado principal
Ba et al. (2024)	China; ECA	77 internos de medicina	Instrucción asistida por IA / MTE	Evaluar la efectividad de la enseñanza asistida por IA frente a los MTE.	La incorporación de IA no afectó la adquisición de conocimientos teóricos, pero favoreció las habilidades clínicas evaluadas: 66,6 % frente a 36,8 % (p = 0,032).
Borg et al. (2026)	Suecia; Estudio cuasiexperimental no aleatorizado	115 estudiantes de medicina	Retroalimentación generada por IA tras interacciones con PV / Interacciones con PV sin retroalimentación de IA	Evaluar el efecto de la retroalimentación con IA sobre el desempeño clínico.	La retroalimentación con IA mejoró el desempeño clínico: 7,39 ± 0,86 frente a 6,68 ± 1,04. La diferencia fue de 0,70 puntos y se estimó entre 0,35 y 1,06 puntos con un 95 % de confianza (p < 0,001).
Brügge et al. (2024)	Alemania; ECA	21 estudiantes de medicina	Paciente simulado mediante IA basado en un LLM y retroalimentación estructurada / Misma simulación sin retroalimentación	Determinar si la retroalimentación ofrecida por un LLM mejora la toma de decisiones clínicas.	Las simulaciones con LLM apoyaron el entrenamiento de la toma de decisiones clínicas de los estudiantes, en especial cuando se combinaron con retroalimentación estructurada: 3,60 ± 0,13 frente a 3,02 ± 0,12 (p = 0,049).
Çiçek et al. (2025)	Turquía; ECA	129 estudiantes de medicina (115 completaron)	Retroalimentación generada por IA / Retroalimentación escrita por expertos.	Evaluar la efectividad de la retroalimentación generada por IA para mejorar las habilidades de RC de los estudiantes.	No se evidenciaron diferencias significativas en el rendimiento general: 74,7 ± 15,1 con ChatGPT frente a 78,5 ± 20,6 con retroalimentación experta (p = 0,26); a los 10 días, 76,0 ± 14,5 frente a 78,0 ± 21,2 (p = 0,57). En los casos complejos, la retroalimentación experta fue superior (p < 0,001).
Dang et al. (2026)	Vietnam; ECA	67 estudiantes de medicina	PEV impulsados por IA / PE convencionales	Evaluar la eficacia de los PEV impulsados por IA en la enseñanza clínica.	Ambos métodos mejoraron el desempeño de los estudiantes de medicina, sin una ventaja estadísticamente significativa de la IA: 6,5 ± 1,5 frente a 6,3 ± 1,7 (p = 0,543). La satisfacción estudiantil también fue comparable entre ambas modalidades de enseñanza.
Espírito Santo et al. (2026)	Brasil; Estudio cuantitativo, exploratorio y multiinstitucional	40 estudiantes de medicina	Aprendizaje asistido por IA (ChatGPT) / Desempeño previo sin apoyo de IA	Evaluar la efectividad de la IA (ChatGPT) como una herramienta educativa de apoyo.	El aprendizaje asistido por IA se asoció con una mejora en el desempeño de los estudiantes de medicina, con un incremento de 92 ± 0,7 % a 99 ± 0,1 % (p < 0,0001).
Gan et al. (2024)	China; ECA	129 estudiantes (110 completaron)	Aprendizaje asistido por IA (ChatGPT) / Aprendizaje mediante recursos digitales convencionales	Evaluar el efecto del aprendizaje asistido por IA en la educación médica.	El grupo con IA obtuvo una puntuación superior en el examen que el grupo control con aprendizaje mediante recursos digitales convencionales: 138,46 frente a 130,80 (p = 0,04), pero no presentó ventaja en el análisis de casos (p = 0,16).
Gao et al. (2026)	China; ECA	40 estudiantes de medicina	Paciente simulado mediante IA / MTE	Evaluar la eficacia de un modelo de paciente simulado basado en IA en comparación con MTE.	Las simulaciones con IA mejoraron el desempeño en los estudiantes de medicina, constituyendo una herramienta pedagógica eficaz: 93,07 ± 3,93 % frente a 88,98 ± 4,05 % (p = 0,0025).
Gu et al. (2026)	China; Estudio piloto educativo aleatorizado	74 estudiantes de medicina	ABP asistido por IA / MTE	Evaluar la efectividad del ABP asistido por IA en estudiantes.	El grupo con IA obtuvo un mejor desempeño que el grupo tradicional: 6,32 ± 1,23 frente a 5,43 ± 1,39 (p = 0,005). El ABP asistido por IA mostró potencial para fortalecer el RC y otras habilidades clínicas evaluadas.
Kafke et al. (2023)	Alemania; ECA	166 estudiantes de medicina	SADC basado en IA / Métodos convencionales de búsqueda	Evaluar la utilidad de un SADC sobre la precisión diagnóstica y la confianza de los estudiantes.	La precisión diagnóstica fue comparable entre ambos métodos y no presentó diferencias estadísticamente significativas (p = 0,197).
Kirpalani y Wang (2026)	Canadá; Estudio pretest-postest de un solo grupo	57 estudiantes de medicina	Recomendaciones generadas mediante IA para la resolución de preguntas médicas / Resolución de preguntas médicas sin apoyo de IA	Evaluar la influencia de la IA en la toma de decisiones de estudiantes.	Las recomendaciones generadas mediante IA aumentaron de manera moderada la precisión, de 46,9 % a 50,7 % (p = 0,0027), pero hicieron que el 8,3 % cambiara una respuesta correcta por una incorrecta.
Li et al. (2025)	China; Estudio comparativo	79 estudiantes de medicina	ABC asistido por IA generativa / ABC mediante MTE	Evaluar el efecto de la IA generativa en el ABC.	El uso de IA mejoró el análisis de casos y las calificaciones académicas: 77,3 ± 4,3 frente a 66,5 ± 5,4 (p < 0,05), además de reducir el tiempo destinado a las actividades de aprendizaje.
McCarrick et al. (2025)	Irlanda; ECA	90 estudiantes de medicina	Paciente simulado mediante IA / MTE	Evaluar el efecto de la simulación con IA en el desarrollo estudiantil.	El grupo con IA presentó una mejoría significativa, con un efecto pequeño a moderado (d de Cohen = 0,37), mientras que la mejoría del grupo tradicional fue menor (d = 0,19).
Montagna et al. (2025)	Italia; Estudio comparativo con asignación aleatoria por grupos	16 estudiantes de medicina	Resolución de casos clínicos mediante SADC asistida por IA / Resolución de casos mediante SADC tradicionales	Evaluar el desempeño de los estudiantes en la resolución de casos mediante diferentes SADC.	Las GPC alcanzaron las puntuaciones medias más altas de precisión e integridad, sin una diferencia global significativa en integridad: 3,8 con ChatGPT frente a 4,0 con GPC (p = 0,49); exactitud global: p = 0,02.
Niu et al. (2026)	China; Estudio controlado con asignación aleatoria	177 estudiantes de medicina	Modelo de enseñanza asistido por IA / MTE	Evaluar la efectividad de un modelo de enseñanza asistido por IA sobre el aprendizaje y toma de decisiones clínicas.	El modelo asistido por IA mejoró el aprendizaje y la precisión en la toma de decisiones clínicas: 81,2 ± 8,1 % frente a 61,6 ± 10,3 % (p < 0,001).
Roemer et al. (2026)	Alemania; ECA	68 estudiantes de medicina	Resolución de casos clínicos escritos con LLM combinado con MTE / Resolución de casos clínicos escritos con MTE	Investigar el impacto de un LLM sobre el desempeño diagnóstico.	Los estudiantes de medicina que usaron un LLM lograron mayor precisión diagnóstica que los que usaron MTE: 77,5 % frente a 32,4 % (p < 0,001).
Rother et al. (2026)	Alemania; Ensayo cruzado aleatorizado	130 estudiantes de medicina	Retroalimentación personalizada generada por IA / Retroalimentación general proporcionada por expertos	Evaluar el efecto de la retroalimentación con IA sobre el RC.	La retroalimentación personalizada generada por IA no produjo un mayor efecto de aprendizaje: 64,4 ± 20,9 % frente a 62,4 ± 22,2 % (p = 0,454).

Saloojee et al. (2026)	Sudáfrica; ECA	73 estudiantes de medicina	Uso de IA generativa por estudiantes durante la evaluación clínica / Evaluación clínica convencional	Evaluar el efecto de la IA sobre el desempeño clínico en pacientes reales.	La utilización de IA durante la evaluación clínica no produjo una mejora en el desempeño clínico global de los estudiantes frente a la evaluación convencional: 66,7 ± 9,9 frente a 68,2 ± 7,8 (p = 0,50).
Shen et al. (2026)	China; ECA	80 estudiantes de medicina	PEV impulsados por IA combinados con capacitación tradicional mediante PE / Capacitación tradicional mediante PE	Evaluar la efectividad educativa cuando se combina PEV con PE.	La combinación de PEV y PE se asoció con un mejor desempeño en RC frente al método convencional: 1,3 ± 0,7 frente a 0,3 ± 0,5 (p < 0,01). Además, favoreció el desempeño clínico, la comunicación, la atención al paciente y la adquisición de conocimientos.
Sindhvananda et al. (2026)	Tailandia; Estudio cuasiexperimental de métodos mixtos	25 estudiantes de medicina	Paciente simulado impulsado por IA con retroalimentación automatizada / Desempeño previo sin apoyo de IA	Evaluar la efectividad de pacientes simulados impulsados por IA para mejorar el desempeño de los estudiantes.	La puntuación de RC aumentó de 28,2 ± 5,2 a 37,0 ± 2,2 (p < 0,001), con un tamaño del efecto grande (d de Cohen = 1,81). También fortaleció la confianza y benefició a estudiantes que presentaban menor desempeño inicial.
Tee et al. (2026)	Taiwán; Estudio observacional prospectivo	26 estudiantes de medicina	Entrenamiento asistido por IA para la interpretación de estudios de imagen / Interpretación radiológica convencional	Evaluar el efecto de la asistencia mediante IA sobre el desempeño diagnóstico.	La asistencia mediante IA mejoró el desempeño diagnóstico: la precisión aumentó 26,7 % (p < 0,01) y mantuvo una mejoría de 13,3 % después de retirarla (p = 0,010).
Teng et al. (2026)	China; ECA	111 estudiantes de medicina	Explicaciones clínicas generadas mediante IA / Evaluación clínica convencional	Evaluar el efecto de las explicaciones generadas mediante IA sobre la precisión diagnóstica y la confianza clínica.	La información engañosa generada por IA redujo la precisión diagnóstica de los estudiantes a 9,2 %, frente a 21,0 % en el control (p < 0,001); la información correcta generada por IA no produjo una mejoría significativa.
Wang et al. (2025)	China; ECA	56 estudiantes de medicina	Paciente simulado mediante IA generativa (GPT) / MTE	Evaluar la efectividad del uso de pacientes simulados por IA en el desarrollo de habilidades clínicas.	El grupo con simulación mediante IA tuvo un mejor desempeño que el grupo tradicional: 86,79 ± 5,46 frente a 73,64 ± 4,76 (p < 0,001). La simulación favoreció la comunicación, la retroalimentación y el aprendizaje autodirigido.
Xie et al. (2025)	China; Estudio comparativo no aleatorizado	97 estudiantes de medicina	Entrenamiento diagnóstico mediante imágenes generadas por IA / Enseñanza mediante casos clínicos e imágenes médicas reales	Evaluar la eficacia de imágenes generadas por IA para mejorar la precisión diagnóstica.	Las imágenes generadas por IA mejoraron la precisión diagnóstica, con resultados comparables a los MTE: la precisión aumentó de 42,68 % a 71,27 %, mejoría: 28,43 %, sin diferencias entre métodos.
J. Xu et al. (2026)	China; ECA	80 estudiantes de medicina	Plataforma multimodal basada en IA combinada con MTE / MTE	Evaluar la eficacia de la intervención con IA para mejorar los resultados de aprendizaje clínico.	El grupo con IA mejoró notablemente las competencias clínicas generales: 85,7 ± 5,8 frente a 78,3 ± 7,1 (p < 0,001). La plataforma también favoreció la integración del conocimiento y el aprendizaje autodirigido.
L. Xu et al. (2025)	China; ECA	60 estudiantes de medicina	Entrenamiento con PEV impulsados por IA / MTE	Evaluar la eficacia de los PEV para mejorar el pensamiento clínico.	El entrenamiento con PEV impulsados por IA produjo un mejor desempeño en pensamiento clínico frente a MTE: 42,60 ± 9,56 frente a 31,63 ± 7,24 (p < 0,001), con un tamaño del efecto grande (d de Cohen = 1,09).
Yamamoto et al. (2024)	Japón; Estudio controlado no aleatorizado	145 estudiantes de medicina	Pacientes simulados por IA / MTE con PE	Evaluar la eficacia de pacientes simulados por IA para mejorar habilidades clínicas.	El entrenamiento con pacientes simulados por IA mejoró el desempeño clínico: 28,1 ± 1,6 frente a 27,1 ± 2,2 (p = 0,01).
Yanagita et al. (2025)	Japón; ECA	42 estudiantes de medicina	SADC basado en IA / Resolución diagnóstica sin sistema de apoyo	Evaluar el efecto de un sistema de apoyo diagnóstico sobre la precisión diagnóstica.	La precisión diagnóstica global fue de 64,1 % con el SADC y de 50,5 % en el grupo control. Con calificadores semánticos apropiados, la precisión fue mayor en el grupo que utilizó el sistema (p < 0,001).
Yang et al. (2026)	China; ECA	120 internos de medicina	Modelo de enseñanza basado en ABP y ABC, asistido por IA / MTE	Evaluar la eficacia de un modelo de enseñanza asistido por IA para fortalecer el RC.	El modelo asistido por IA mejoró las habilidades clínicas y el RC: 4,54 ± 0,21 frente a 4,13 ± 0,21 (p < 0,0001). Además de estimular el interés por aprender y la capacidad de aprendizaje autónomo.
Yu et al. (2026)	China; Estudio de cohorte histórica	190 estudiantes de medicina	Enseñanza entre pares asistida por IA / Enseñanza tradicional entre pares	Evaluar el efecto de la enseñanza entre pares asistida por IA en el RC.	El grupo asistido por IA superó al comparador en 2,22 puntos en RC; la diferencia se estimó entre 1,18 y 3,25 puntos con un 95 % de confianza (p < 0,001; d de Cohen = 0,61).

Nota. ABC: aprendizaje basado en casos; ABP: aprendizaje basado en problemas; ECA: ensayo controlado aleatorizado; GPC: guía de práctica clínica; GPT: *Generative Pre-trained Transformer*; IA: inteligencia artificial; LLM: modelos de lenguaje de gran escala; MTE: métodos tradicionales de enseñanza; PE: pacientes estandarizados; PEV: pacientes estandarizados virtuales; PV: pacientes virtuales; RC: razonamiento clínico; SADC: sistema de apoyo a la decisión clínica.

Fuente: Elaboración propia.

La dirección predominante de los resultados fue favorable. Diecinueve estudios registraron mejorías en uno o varios componentes del razonamiento clínico, desde la obtención y organización de información hasta la formulación diagnóstica, el criterio clínico, la toma de decisiones y la planificación terapéutica (Ba et al., 2024; Borg et al., 2026; Brügge et al., 2024; Espírito Santo et al., 2026; Gao et al., 2026; Gu et al., 2026; Li et al., 2025; McCarrick et al., 2025; Niu et al., 2026; Roemer et al., 2026; Shen et al., 2026; Sindhvananda et al., 2026; Tee et al., 2026; Wang et al., 2025; J. Xu et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024; Yang et al., 2026; Yu et al., 2026). El alcance varió entre las intervenciones, debido a que algunas favorecieron varias dimensiones y otras produjeron beneficios más específicos. En siete estudios, el desempeño alcanzado con IA fue comparable al obtenido mediante enseñanza convencional, retroalimentación experta, pacientes estandarizados o recursos clínicos tradicionales. Estos resultados respaldan su utilidad como apoyo educativo, pero no demuestran una superioridad constante frente al comparador (Çiçek et al., 2025; Dang et al., 2026; Kafke et al., 2023; Montagna et al., 2025; Rother et al., 2026; Saloojee et al., 2026; Xie et al., 2025).

Tres investigaciones mostraron efectos mixtos o condicionados por la actividad, la calidad de la información proporcionada y la manera de utilizar la herramienta (Gan et al., 2024; Kirpalani y Wang, 2026; Yanagita et al., 2025). El efecto predominantemente desfavorable se identificó ante la exposición a información engañosa, que redujo la precisión diagnóstica y alteró la calibración de la confianza (Teng et al., 2026). La Tabla 2 presenta esta clasificación.

Tabla 2. *Patrones de efecto observados después de incorporar IA a la formación médica.*

Dirección del efecto	Estudios	Alcance e interpretación
Predominantemente favorable	Ba et al. (2024); Borg et al. (2026); Brügge et al. (2024); Espírito Santo et al. (2026); Gao et al. (2026); Gu et al. (2026); Li et al. (2025); McCarrick et al. (2025); Niu et al. (2026); Roemer et al. (2026); Shen et al. (2026); Sindhvananda et al. (2026); Tee et al. (2026); Wang et al. (2025); J. Xu et al. (2026); L. Xu et al. (2025); Yamamoto et al. (2024); Yang et al. (2026); Yu et al. (2026)	La IA se asoció con mejorías en uno o varios componentes del RC. El alcance comprendió la obtención y organización de información, la formulación de hipótesis, el criterio clínico, el diagnóstico, la toma de decisiones y la planificación terapéutica.
Comparable frente al comparador	Çiçek et al. (2025); Dang et al. (2026); Kafke et al. (2023); Montagna et al. (2025); Rother et al. (2026); Saloojee et al. (2026); Xie et al. (2025)	El desempeño alcanzado con IA fue semejante al obtenido mediante enseñanza convencional, retroalimentación experta, pacientes estandarizados o recursos clínicos tradicionales, sin presentar una ventaja consistente.
Mixto o condicionado	Gan et al. (2024); Kirpalani y Wang (2026); Yanagita et al. (2025)	Los beneficios dependieron de la actividad, la calidad de la información o la manera de utilizar la herramienta. También se observaron resultados neutrales, errores inducidos o indicios de sesgo de automatización.
Predominantemente desfavorable	Teng et al. (2026)	La exposición a información engañosa generada mediante IA redujo la precisión diagnóstica y produjo una confianza que no correspondía con la calidad de las respuestas.

Nota. La clasificación representa el criterio interpretativo de los autores de la revisión y no constituye una estimación estadística conjunta. IA: inteligencia artificial; RC: razonamiento clínico.

Fuente: Elaboración propia.

La valoración metodológica identificó un estudio con riesgo bajo, 13 con riesgo moderado y 16 con riesgo alto para la atribución causal de los resultados. Las preocupaciones más frecuentes fueron la ausencia de asignación aleatoria, la falta de medición basal, las intervenciones combinadas, las muestras pequeñas, los controles históricos y las evaluaciones realizadas inmediatamente después del entrenamiento. Estas condiciones no invalidan los resultados, pero limitan la confianza con la que puede interpretarse la relación entre las intervenciones y los efectos observados.Los estudios clasificados con riesgo alto conservaron utilidad descriptiva y permitieron reconocer aplicaciones educativas, tendencias y posibles riesgos de la IA. Sin embargo, sus resultados no recibieron el mismo peso interpretativo que los ensayos aleatorizados

con comparadores equivalentes y evaluaciones independientes. La Tabla 3 presenta el instrumento aplicado, la valoración global y el fundamento principal de cada decisión.

Tabla 3. Evaluación del riesgo de sesgo de los estudios incluidos

Referencia	Instrumento	Valoración global	Fundamento principal
Ba et al. (2024)	RoB 2	Moderado	Existieron asignación aleatoria, casos equivalentes y evaluación estructurada; la ausencia de medición basal y la información limitada sobre la secuencia de asignación reducen la certeza.
Borg et al. (2026)	ROBINS-I	Alto	La asignación dependió de la sede hospitalaria y no se realizó una evaluación basal que permitiera comprobar la equivalencia clínica de los grupos.
Brügge et al. (2024)	RoB 2	Moderado	La aleatorización y la evaluación independiente fortalecen el estudio; la muestra reducida, las exclusiones técnicas y las mediciones repetidas limitan su precisión.
Çiçek et al. (2025)	RoB 2	Moderado	Los participantes desconocían el origen de la retroalimentación, pero existieron pérdidas durante el seguimiento y el estudio se realizó en una sola institución.
Dang et al. (2026)	RoB 2	Moderado	La asignación por clases y la evaluación mediante ECOE aportan una comparación pertinente, aunque se reconoce que la muestra fue pequeña y la aleatorización individual no estuvo presente.
Espírito Santo et al. (2026)	JBI	Alto	No existió grupo control, se utilizó el mismo caso antes y después y la segunda actividad se realizó inmediatamente después de recibir retroalimentación.
Gan et al. (2024)	RoB 2	Alto	Se registraron pérdidas durante el seguimiento, el análisis incluyó únicamente a quienes completaron el estudio y aparecieron discrepancias entre los resultados comunicados.
Gao et al. (2026)	RoB 2	Alto	Aunque la asignación fue aleatoria, no se efectuó evaluación basal, los evaluadores conocían la intervención y la muestra fue pequeña y procedió de una sola institución.
Gu et al. (2026)	RoB 2	Moderado	La asignación y la evaluación fueron adecuadas, pero la IA se integró con ABP, trabajo grupal y simulación, lo que dificulta separar su efecto específico.
Kafke et al. (2023)	RoB 2	Moderado	El diseño cruzado permitió comparar ambos métodos, aunque la repetición del mismo caso sin periodo de separación pudo influir en la segunda resolución.
Kirpalani y Wang (2026)	JBI	Alto	No existió grupo control y los participantes respondieron las mismas preguntas antes y después de revisar las recomendaciones de ChatGPT.
Li et al. (2025)	ROBINS-I	Alto	Se utilizaron clases preexistentes, no hubo medición basal y se excluyeron estudiantes según su actitud previa hacia la IA.
McCarrick et al. (2025)	RoB 2	Moderado	La asignación por conglomerados, las mediciones antes-después y los examinadores independientes fortalecen el estudio; el grupo experimental recibió práctica adicional.
Montagna et al. (2025)	RoB 2	Alto	La muestra fue reducida, cada condición estuvo representada por un solo equipo y las respuestas grupales impidieron valorar el desempeño individual.
Niu et al. (2026)	RoB 2	Moderado	Se informó asignación mediante tabla de números aleatorios y comparabilidad basal; la ausencia de información sobre ocultamiento y el uso de un instrumento adaptado reducen la certeza.
Roemer et al. (2026)	RoB 2	Moderado	El ensayo fue aleatorizado y empleó resultados diagnósticos objetivos; el carácter abierto y las respuestas repetidas por estudiante requieren cautela.
Rother et al. (2026)	RoB 2	Alto	La participación y utilización de la aplicación fueron voluntarias, el examen final no fue obligatorio y existieron diferencias importantes en la intensidad de uso.
Saloojee et al. (2026)	RoB 2	Moderado	La aleatorización, la evaluación independiente y el análisis por intención de tratar fortalecen el estudio; la utilización variable de ChatGPT, la capacitación incompleta y la ausencia de registros detallados de uso introducen algunas preocupaciones.
Shen et al. (2026)	RoB 2	Moderado	La secuencia aleatoria, el ocultamiento de la asignación hasta completar la inscripción y la evaluación basal, y la ausencia de pérdidas fortalecen el diseño del estudio; el grupo experimental recibió una exposición formativa adicional.
Sindhvananda et al. (2026)	JBI	Alto	No existió grupo control, los evaluadores conocían el momento de evaluación y continuó el estudio autónomo durante la intervención.
Tee et al. (2026)	JBI	Alto	La ausencia de control, la repetición de imágenes y el seguimiento breve impiden diferenciar la contribución de la IA del aprendizaje inmediato.
Teng et al. (2026)	RoB 2	Bajo	La asignación aleatoria, la medición objetiva y el análisis que consideró las respuestas repetidas permitieron una evaluación sólida del efecto y de la desinformación.
Wang et al. (2025)	RoB 2	Moderado	Se aplicaron asignación aleatoria y mediciones antes-después; la muestra pequeña, el carácter unicéntrico y la falta de información sobre el conocimiento de la asignación por parte de los evaluadores reducen la certeza.
Xie et al. (2025)	ROBINS-I	Alto	Los grupos no fueron aleatorizados, la evaluación fue inmediata y la exposición repetida a imágenes pudo favorecer la memoria visual.
J. Xu et al. (2026)	RoB 2	Alto	No se comunicó una medición basal, el grupo experimental recibió práctica adicional y se identificaron inconsistencias en determinados datos del reporte.
L. Xu et al. (2025)	RoB 2	Alto	Parte del desenlace se evaluó mediante la plataforma empleada durante la formación y el comparador no recibió una experiencia completamente equivalente.
Yamamoto et al. (2024)	ROBINS-I	Alto	El grupo control procedió de una cohorte histórica, por lo que las diferencias temporales y curriculares pudieron influir en el resultado.
Yanagita et al. (2025)	RoB 2	Alto	La muestra fue pequeña y el análisis no consideró plenamente la dependencia entre las respuestas repetidas de cada estudiante.
Yang et al. (2026)	RoB 2	Moderado	La asignación fue aleatoria, pero la intervención combinó IA, ABP y ABC, lo que impide determinar el efecto individual de cada componente.
Yu et al. (2026)	ROBINS-I	Alto	Se compararon promociones de años diferentes y no fue posible controlar completamente la utilización adicional de IA fuera de las actividades programadas.

Nota. La clasificación representa la interpretación metodológica del equipo revisor a partir de la información comunicada en cada publicación y no constituye una calificación absoluta del estudio o de la revista. El instrumento se seleccionó según el diseño y la valoración dependió de las limitaciones identificadas; por esta razón, un mismo instrumento puede presentar diferentes niveles de riesgo. ABC: aprendizaje basado en casos; ABP: aprendizaje basado en problemas; ECOE: examen clínico objetivo estructurado; IA: inteligencia artificial; JBI: lista de verificación del Joanna Briggs Institute para estudios antes-después; RoB 2: herramienta para valorar el riesgo de sesgo en estudios aleatorizados; ROBINS-I: herramienta para valorar el riesgo de sesgo en estudios comparativos no aleatorizados.

Fuente: Elaboración propia.

El análisis conjunto de las Tablas 2 y 3 permitió establecer que la frecuencia de resultados favorables no estuvo acompañada por un nivel equivalente de confianza metodológica. De los 19 estudios predominantemente favorables, nueve presentaron riesgo moderado y diez presentaron riesgo alto. Entre los siete estudios comparables, cuatro fueron clasificados con riesgo moderado y

tres con riesgo alto; las tres investigaciones con efectos mixtos recibieron una valoración de riesgo alto. El único estudio clasificado con riesgo bajo examinó la exposición a información engañosa y documentó un efecto desfavorable sobre la precisión diagnóstica y la calibración de la confianza (Teng et al., 2026). Este hallazgo no contradice los beneficios observados en contextos de información correcta, pero delimita con mayor confianza metodológica un riesgo de seguridad: las recomendaciones incorrectas pueden deteriorar el razonamiento y mantener una confianza injustificada.

Esta distribución impide valorar la contribución educativa de la IA únicamente a partir de la cantidad de intervenciones favorables. Los resultados procedentes de estudios con riesgo moderado o alto respaldan un beneficio potencial, pero limitan la atribución causal y la posibilidad de trasladarlo hacia otros contextos formativos. Los hallazgos comparables y desfavorables delimitan las condiciones en las que la tecnología no aporta una ventaja o puede interferir con el desempeño estudiantil. En respuesta a la pregunta de investigación, las herramientas de IA pueden favorecer el razonamiento clínico de los estudiantes de medicina, pero sus efectos no son uniformes ni garantizan superioridad frente a los métodos convencionales.

El análisis realizado permite precisar que su contribución depende del diseño pedagógico, la complejidad clínica, la calidad de las respuestas, el comparador seleccionado y la capacidad del estudiante para examinarlas críticamente. El análisis de los estudios incluidos determinó que la IA puede favorecer el razonamiento clínico de los estudiantes de medicina cuando forma parte de experiencias educativas estructuradas. La amplitud del efecto dependió de la actividad, el acompañamiento docente, la calidad de la información y la participación del estudiante. Su incorporación aislada no aseguró una mejor interpretación clínica ni decisiones superiores a las obtenidas con métodos convencionales.

Una coincidencia entre las investigaciones favorables fue la participación activa del estudiante en actividades clínicas definidas. Los pacientes simulados, las plataformas multimodales y las estrategias asistidas por IA exigieron entrevistar, seleccionar información, formular hipótesis, establecer diagnósticos o justificar planes de manejo o tratamiento. La tecnología aportó casos, recursos y retroalimentación; el aprendizaje dependió del trabajo intelectual realizado durante la resolución clínica (Ba et al., 2024; Borg et al., 2026; Brügge et al., 2024; Espírito Santo et al., 2026; Gao et al., 2026; Gu et al., 2026; Li et al., 2025; McCarrick et al., 2025; Niu et al., 2026; Roemer et al., 2026; Shen et al., 2026; Sindhvananda et al., 2026; Wang et al., 2025; J. Xu et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024; Yang et al., 2026; Yu et al., 2026).

La práctica reiterada y la retroalimentación inmediata también estuvieron presentes en intervenciones desarrolladas con herramientas diferentes. Repetir entrevistas, reconocer omisiones y revisar decisiones permitió un aprendizaje progresivo dentro de entornos controlados. Esta coincidencia sugiere que parte del beneficio puede relacionarse con la práctica deliberada y la organización pedagógica de la experiencia, además de la asistencia proporcionada por el sistema tecnológico (Borg et al., 2026; Brügge et al., 2024; Gao et al., 2026; McCarrick et al., 2025; Shen et al., 2026;

Sindhvananda et al., 2026; Wang et al., 2025; L. Xu et al., 2025; Yamamoto et al., 2024). Las diferencias se relacionaron con los componentes del razonamiento clínico beneficiados por cada intervención. En algunos estudios, la mejoría se concentró en sus etapas iniciales: anamnesis, identificación de información pertinente, organización de los datos y secuencia de las preguntas. Estas habilidades permiten construir una representación clínica del problema, aunque su progreso no se acompañó de avances equivalentes en el criterio clínico o la competencia global (Borg et al., 2026; Espírito Santo et al., 2026; Gao et al., 2026; McCarrick et al., 2025; Sindhvananda et al., 2026; Yamamoto et al., 2024).

Otras intervenciones extendieron sus efectos al diagnóstico diferencial, el desempeño diagnóstico, el criterio clínico, la toma de decisiones y la planificación terapéutica. Esta diferencia confirma que el alcance educativo varía según la actividad propuesta y el desenlace evaluado. La obtención ordenada de información contribuye al razonamiento clínico, pero no asegura por sí misma una decisión apropiada (Ba et al., 2024; Brügge et al., 2024; Gu et al., 2026; Li et al., 2025; Niu et al., 2026; Roemer et al., 2026; Tee et al., 2026; Wang et al., 2025; J. Xu et al., 2026; L. Xu et al., 2025; Yang et al., 2026; Yu et al., 2026).

Los estudios con desempeño comparable delimitaron con mayor precisión el aporte educativo de la IA. Los pacientes virtuales alcanzaron niveles semejantes a los pacientes estandarizados; la retroalimentación automatizada no superó globalmente la elaborada por expertos; y los sistemas de apoyo diagnóstico obtuvieron valores próximos a los recursos tradicionales. Durante encuentros con pacientes reales tampoco se documentó una ventaja clínica significativa (Çiçek et al., 2025; Dang et al., 2026; Kafke et al., 2023; Montagna et al., 2025; Rother et al., 2026; Saloojee et al., 2026; Xie et al., 2025).

Los resultados comparables frente a los distintos comparadores respaldan el uso complementario de estas herramientas, pero no justifican la sustitución de los métodos convencionales. Su valor educativo puede encontrarse en la ampliación de las oportunidades de práctica, la disponibilidad permanente y la posibilidad de ofrecer escenarios diversos. La superioridad académica o clínica requiere una demostración específica para cada herramienta, población y actividad evaluada (Çiçek et al., 2025; Dang et al., 2026; Kafke et al., 2023; Montagna et al., 2025; Rother et al., 2026; Saloojee et al., 2026; Xie et al., 2025).

Los efectos mixtos precisaron las condiciones que modifican la utilidad de la herramienta. Un sistema de apoyo diagnóstico alcanzó mayor eficacia cuando el estudiante introdujo términos clínicos pertinentes. Otras intervenciones mejoraron parte del desempeño académico, pero no el análisis de casos, y algunas recomendaciones corrigieron respuestas erróneas a la vez que indujeron nuevos errores. El beneficio dependió del conocimiento previo y de la valoración crítica de cada recomendación (Gan et al., 2024; Kirpalani & Wang, 2026; Yanagita et al., 2025). El principal riesgo se relacionó con la aceptación de información incorrecta generada por IA. Las explicaciones engañosas redujeron la precisión diagnóstica sin disminuir la confianza; algunos estudiantes permanecieron seguros de respuestas equivocadas. Otras recomendaciones

llevaron a modificar decisiones inicialmente correctas o a seguir al sistema durante la interpretación de imágenes. Este comportamiento representa sesgo de automatización: confiar en la herramienta por encima del análisis clínico propio (Kirpalani & Wang, 2026; Tee et al., 2026; Teng et al., 2026).

La relación entre las tres tablas permite interpretar los efectos de acuerdo con su contexto metodológico. La Tabla 1 presenta las intervenciones, los comparadores y los resultados principales; la Tabla 2 organiza la dirección de los efectos; y la Tabla 3 expone el riesgo de sesgo. Los beneficios procedentes de intervenciones combinadas o cohortes históricas requieren mayor cautela, y la desinformación adquiere especial relevancia al proceder del estudio con menor riesgo de sesgo.

Para los docentes, la IA debe incorporarse dentro de una secuencia pedagógica con objetivos, instrucciones y criterios de evaluación definidos. Puede ampliar la disponibilidad de casos, facilitar prácticas repetidas y ofrecer retroalimentación inicial antes de la revisión académica. Su utilidad aumenta cuando el tiempo, los pacientes estandarizados o los escenarios de simulación son limitados, siempre bajo supervisión docente (Borg et al., 2026; Brügge et al., 2024; Dang et al., 2026; Gao et al., 2026; Gu et al., 2026; McCarrick et al., 2025; Shen et al., 2026; Sindhvananda et al., 2026; Wang et al., 2025; J. Xu et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024). La integración curricular debe incluir el manejo técnico de la herramienta y el examen crítico de sus respuestas. El estudiante necesita formular consultas precisas, reconocer información insuficiente, contrastar recomendaciones con fuentes médicas validadas y justificar cada decisión. La respuesta generada debe funcionar como material de análisis, acompañada de una argumentación clínica propia que preserve la autonomía intelectual del futuro médico (Kirpalani & Wang, 2026; Montagna et al., 2025; Tee et al., 2026; Teng et al., 2026; Yanagita et al., 2025).

La supervisión docente conserva una función esencial en la selección de casos, la verificación del contenido y la evaluación de la argumentación. La retroalimentación experta mantuvo ventajas ante casos que exigían integrar información clínica menos evidente (Çiçek et al., 2025; Rother et al., 2026). En la comparación entre sistemas de apoyo diagnóstico, las guías de práctica clínica obtuvieron las puntuaciones medias más altas de exactitud e integridad. La diferencia de exactitud se concentró en la comparación entre las guías de práctica clínica y los repositorios en línea, mientras que la integridad no mostró diferencias globales significativas (Montagna et al., 2025). La IA ofrece disponibilidad y acceso oportuno; el docente proporciona profundidad, contextualización y responsabilidad académica.

La aplicación próxima a la atención asistencial exige especial prudencia. El desempeño alcanzado en simulación no asegura resultados equivalentes durante el contacto con pacientes reales, donde intervienen comunicación no verbal, incertidumbre, presión temporal y responsabilidad profesional. Estas herramientas pueden contribuir a la preparación previa, sin desplazar la entrevista supervisada, la exploración física, el vínculo terapéutico ni la toma de decisiones compartida (Dang et al., 2026; Gao et al., 2026; Saloojee et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024). Las

investigaciones futuras deben separar el efecto de la IA del producido por otras estrategias educativas. Cuando el grupo experimental recibe más casos, mayor tiempo de práctica, ABP, aprendizaje basado en casos, enseñanza entre pares o retroalimentación adicional, la diferencia final no puede atribuirse exclusivamente a la tecnología. Los comparadores activos con exposición formativa equivalente y los diseños factoriales permitirían estimar la contribución independiente de cada componente (Borg et al., 2026; Gu et al., 2026; McCarrick et al., 2025; Shen et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024; Yang et al., 2026; Yu et al., 2026).

También se requieren estudios multicéntricos, muestras amplias, mediciones basales y evaluadores independientes que desconozcan la asignación. La aleatorización debe acompañarse de ocultamiento, análisis de pérdidas y procedimientos estadísticos apropiados para observaciones repetidas o agrupadas. El registro prospectivo y la definición previa de los desenlaces mejorarían la transparencia y facilitarían la comparación entre investigaciones.

La permanencia de los efectos todavía no puede establecerse con precisión. Las evaluaciones diferidas deberían medir retención, transferencia hacia casos nuevos y desempeño sin asistencia tecnológica. La incorporación de pacientes estandarizados independientes y evaluaciones clínicas, cuando sean éticamente pertinentes, permitiría diferenciar una mejoría inmediata en la actividad practicada de un aprendizaje aplicable a la atención (McCarrick et al., 2025; Rother et al., 2026; Saloojee et al., 2026; Sindhvananda et al., 2026; Tee et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024; Yu et al., 2026).El proceso de interacción entre el estudiante y la IA requiere un análisis específico. Registrar las instrucciones, consultas, respuestas generadas, correcciones y la decisión final permitiría reconocer conductas asociadas con aprendizaje, error o dependencia. La versión del modelo, sus parámetros de uso y las modificaciones docentes también deben documentarse para favorecer la reproducibilidad de los hallazgos (Çiçek et al., 2025; Kirpalani & Wang, 2026; Montagna et al., 2025; Rother et al., 2026; Saloojee et al., 2026; Tee et al., 2026; Teng et al., 2026; Yanagita et al., 2025).

La seguridad cognitiva debe ocupar una posición central entre los desenlaces. Las investigaciones deberían incorporar información correcta, incompleta y engañosa para evaluar la detección de errores, el rechazo de recomendaciones improcedentes y la calibración de la confianza. También conviene examinar la persistencia del sesgo de automatización después de retirar la herramienta y su influencia en decisiones posteriores (Kirpalani y Wang, 2026; Tee et al., 2026; Teng et al., 2026).

La diversidad de instrumentos dificulta las comparaciones directas y limita la posibilidad de realizar un metaanálisis. Un conjunto común de desenlaces debería diferenciar la obtención de información, la formulación de hipótesis, el diagnóstico diferencial, el criterio clínico, la toma de decisiones y la planificación terapéutica. Las evaluaciones objetivas y comparables permitirían precisar qué dimensión cambia, durante cuánto tiempo y con qué relevancia clínica. La heterogeneidad de las intervenciones, los comparadores y los desenlaces constituye la principal limitación de la evidencia. Los estudios examinaron actividades diferentes, desde anamnesis y retroalimentación hasta apoyo

diagnóstico, análisis de casos e interpretación de imágenes. Una estimación estadística única habría tratado como equivalentes componentes y contextos distintos; la integración narrativa conservó su significado clínico y educativo.

La confianza en la evidencia está limitada por el predominio de estudios con riesgo moderado o alto. Las muestras pequeñas, los diseños unicéntricos, la ausencia de medición basal, los controles históricos, las intervenciones combinadas y las evaluaciones inmediatas dificultan la atribución causal. La concentración de investigaciones en instituciones y países específicos también restringe su aplicación en programas con currículos, recursos y culturas educativas diferentes (Borg et al., 2026; Espírito Santo et al., 2026; Gan et al., 2024; Gao et al., 2026; Li et al., 2025; McCarrick et al., 2025; Montagna et al., 2025; Sindhvananda et al., 2026; Xie et al., 2025; J. Xu et al., 2026; L. Xu et al., 2025; Yamamoto et al., 2024; Yu et al., 2026).

La revisión presenta limitaciones propias. Aunque los criterios admitieron publicaciones en inglés o español, todos los estudios incluidos estuvieron escritos en inglés y pertenecieron al periodo comprendido entre enero de 2022 y julio de 2026. El protocolo no contó con registro prospectivo, y la heterogeneidad impidió realizar análisis formales de sensibilidad, evaluar estadísticamente el sesgo de publicación y calificar de manera formal la certeza global. La búsqueda estructurada se concentró en PubMed y Scopus, bases de amplia cobertura y reconocimiento académico; aun así, es posible que no se hayan recuperado estudios indexados exclusivamente en otras fuentes. El componente complementario de la búsqueda presentó menor reproducibilidad que las consultas estructuradas, debido a que las exploraciones iniciales en sitios editoriales no generaron archivos de exportación equivalentes a los de PubMed y Scopus. Para reducir esta limitación se documentaron los títulos, autores, DOI y sitios de recuperación de los nueve informes incorporados mediante otros métodos. Las ecuaciones definitivas, las fechas, los totales brutos y el procedimiento de deduplicación quedaron registrados, lo que permite reproducir de manera independiente la búsqueda estructurada.

La clasificación de los efectos y del riesgo de sesgo requirió juicio interpretativo por parte del equipo revisor. La matriz común, la lectura completa de los artículos y la comprobación por un segundo revisor redujeron la influencia del criterio individual, pero no suprimieron el componente interpretativo inherente a estas valoraciones. La inclusión de resultados favorables, comparables, mixtos y desfavorables permitió evitar una lectura centrada exclusivamente en los beneficios tecnológicos.

Conclusiones

Esta revisión determinó que la IA puede favorecer diversos componentes del razonamiento clínico en estudiantes de medicina. Los efectos favorables se concentraron en actividades estructuradas, con participación activa, práctica reiterada y retroalimentación, y comprendieron la obtención y organización de información, la formulación diagnóstica, el criterio clínico, la toma de decisiones y la planificación terapéutica. La evidencia no permite afirmar que la IA supere de forma uniforme a la enseñanza convencional, la retroalimentación experta o los pacientes estandarizados. Los efectos dependieron del diseño pedagógico, el comparador, la actividad

clínica, la calidad de la información y el conocimiento previo del estudiante. El predominio de estudios con riesgo moderado o alto limita la atribución causal, y la información incorrecta puede deteriorar decisiones acertadas y mantener una confianza injustificada. La IA debe incorporarse como complemento de la formación médica, con supervisión docente y verificación sistemática de sus respuestas. Su integración debe fortalecer la argumentación, el contraste de hipótesis, la detección de errores y la autonomía intelectual. El valor educativo reside en emplearla para profundizar el razonamiento del estudiante y preservar la responsabilidad clínica que sustenta cada decisión.

Referencias bibliográficas

Ba, H., Zhang, L., & Yi, Z. (2024). Enhancing clinical skills in pediatric trainees: A comparative study of ChatGPT-assisted and traditional teaching methods. *BMC Medical Education, 24*(1), 558. https://doi.org/10.1186/s12909-024-05565-1

Barker, T. H., Habibi, N., Aromataris, E., Stone, J. C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S., & Munn, Z. (2024). The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. *JBI Evidence Synthesis, 22*(3), 378–388. https://doi.org/10.11124/JBIES-23-00268

Borg, A., Schiött, J., Ivegren, W., Gentline, C., Huss, V., Hugelius, A. M., Jobs, B., Ruiz, M., Edelbring, S., Georg, C., Skantze, G., & Parodis, I. (2026). AI-generated feedback following social robotic virtual patient interactions and medical student performance: Nonrandomized quasi-experimental study. *JMIR Medical Education, 12*, e90368. https://doi.org/10.2196/90368

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M. H., & Darici, D. (2024). Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Medical Education, 24*(1), 1391. https://doi.org/10.1186/s12909-024-06399-7

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ, 368*, l6890. https://doi.org/10.1136/bmj.l6890

Çiçek, F. E., Ülker, M., Özer, M., & Kıyak, Y. S. (2025). ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: A randomized controlled trial. *Postgraduate Medical Journal, 101*(1195), 458–463. https://doi.org/10.1093/postmj/qgae170

Dang, H. N. N., Luong, T. V., Vo, H. T. H., Nguyen, L. T. K., Tran, T. N., Pham, H. T. A., Truong, H. T., Tran, H., Tran, T. T., Hoang, T. A., Doan, T. C., Huynh, Q., & Marwick, T. H. (2026). Artificial intelligence-powered virtual standardized patients in teaching history-taking skills to medical students: A randomized controlled trial. *BMC Medical Education, 26*(1), 984. https://doi.org/10.1186/s12909-026-09305-5

Espírito Santo, D. S., Bezerra, T. L., Gomes, M. C. V. L., Fonseca, P. A. S., & Joviano-Santos, J. V. (2026). AI-assisted learning: ChatGPT for anamnesis in women's health education. *Medical Education Online, 31*(1), 2655562. https://doi.org/10.1080/10872981.2026.2655562

Gan, W., Ouyang, J., Li, H., Xue, Z., Zhang, Y., Dong, Q., Huang, J., Zheng, X., & Zhang, Y. (2024). Integrating ChatGPT in orthopedic education for medical undergraduates: Randomized controlled trial. *Journal of Medical Internet Research, 26*, e57037. https://doi.org/10.2196/57037

Gao, L., Zhang, L., Zhang, Y., Wu, M., Xu, J., & Lu, D. (2026). Utilizing generative artificial intelligence to create simulated patient for history taking in gynecology. *BMC Medical Education, 26*(1), 1013. https://doi.org/10.1186/s12909-026-09219-2

Gu, Z., Xi, T., & Zhang, C. (2026). DeepSeek-assisted problem-based learning for glaucoma education in an undergraduate ophthalmology clerkship: A randomized educational pilot study. *Scientific Reports, 16*(1), 24184. https://doi.org/10.1038/s41598-026-58447-5

Kafke, S. D., Kuhlmey, A., Schuster, J., Blüher, S., Czimmeck, C., Zoellick, J. C., & Grosse, P. (2023). Can clinical decision support systems be an asset in medical education? An experimental approach. *BMC Medical Education, 23*(1), 570. https://doi.org/10.1186/s12909-023-04568-8

Kirpalani, A., & Wang, P. Z. T. (2026). Medical student reliance on artificial intelligence in nephrology education. *Journal of Nephrology*. Advance online publication. https://doi.org/10.1093/joneph/aajag033

Li, L., Zhang, W., Zhang, K., Yang, Y., Wang, L., Zuo, L., Sun, Y., & Peng, Q. (2025). The role of generative AI tools in case-based learning and teaching evaluation of medical biochemistry. *BMC Medical Education, 25*(1), 1185. https://doi.org/10.1186/s12909-025-07567-z

McCarrick, C. A., McEntee, P. D., Boland, P. A., Donnelly, S., O'Meara, Y., Heneghan, H., & Cahill, R. A. (2025). A randomized controlled trial of a deep language learning model-based simulation tool for undergraduate medical students in surgery. *Journal of Surgical Education, 82*(9), 103629. https://doi.org/10.1016/j.jsurg.2025.103629

Montagna, M., Chiabrando, F., De Lorenzo, R., & Rovere Querini, P. (2025). Impact of clinical decision support systems on medical students' case-solving performance: Comparison study with a focus group. *JMIR Medical Education, 11*, e55709. https://doi.org/10.2196/55709

Niu, Y., Xu, J., Yang, W., Yuan, H., Liu, Z., Zhang, W., & Liu, L. (2026). Application of an iterative generative AI-augmented teaching model in medical immunology education. *Frontiers in Artificial Intelligence, 9*, 1788369. https://doi.org/10.3389/frai.2026.1788369

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Roemer, A., Schlicker, N., Kernder, A., Albe, B., Hack, J., Hirsch, M., Mayr, A., Kuhn, S., & Knitza, J. (2026). Large language models enhance diagnostic reasoning of medical students in rheumatology: A randomized controlled trial. *BMC Medical Education, 26*(1), 579. https://doi.org/10.1186/s12909-026-09079-w

Rother, J. F., Nissen, L., Laupichler, M. C., Raupach, T., & Sommer, N. (2026). Enhancing clinical reasoning in surgical education: A randomized crossover trial on AI-generated individual feedback for key feature question-based online assignments. *Journal of Surgical Education, 83*(9), 104068. https://doi.org/10.1016/j.jsurg.2026.104068

Saloojee, H., Gramanie, M. C., Mwale, R., Bassett, B. A., Madhi, S. A., & Kalla, I. S. (2026). AI at the bedside: Randomised controlled trial of ChatGPT's impact on student performance in real-patient clinical exams. *Medical Teacher*. Advance online publication. https://doi.org/10.1080/0142159X.2026.2652061

Shen, T., Zhang, R., Wang, H., Li, D., & Han, J. (2026). Virtual patients and standardized patients combined training is associated with improved clinical reasoning among medical

students. *Frontiers in Medicine, 13*, 1825465. https://doi.org/10.3389/fmed.2026.1825465

Sindhvananda, K., Ruengwattanachot, K., Leksuwankun, S., Pongpitakmetha, T., Hiransuthikul, A., Surawattanawong, T., Anukoolwittaya, P., & Panjasriprakarn, P. (2026). AI-powered simulated patients with automated feedback for enhancing headache history-taking skills: A convergent mixed-methods study. *BMC Medical Education, 26*(1), 1151. https://doi.org/10.1186/s12909-026-09511-1

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J. C., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, l4898. https://doi.org/10.1136/bmj.l4898

Tee, Y.-S., Liao, C.-A., Kuo, L.-W., Hsu, C.-P., Wang, C.-C., Fu, C.-Y., Liao, C.-H., & Cheng, C.-T. (2026). Artificial intelligence–assisted training for rib fracture interpretation: A prospective study in undergraduate medical students. *World Journal of Emergency Surgery, 21*(1), 14. https://doi.org/10.1186/s13017-026-00678-y

Teng, D., Tan, L., Cao, Q., Xia, Y., Zhang, N., Li, J., & Zhao, D. (2026). Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. *npj Digital Medicine, 9*(1), 356. https://doi.org/10.1038/s41746-026-02547-z

Wang, Z., Fan, T.-T., Li, M.-L., Zhu, N.-J., & Wang, X.-C. (2025). Feasibility study of using GPT for history-taking training in medical education: A randomized clinical trial. *BMC Medical Education, 25*(1), 1030. https://doi.org/10.1186/s12909-025-07614-9

Xie, W., Yuan, Z., Si, Y., Huang, Z., Li, Y., Wu, F., & Yao, Y.-F. (2025). Enhancing medical students' diagnostic accuracy of infectious keratitis with AI-generated images. *BMC Medical Education, 25*(1), 1027. https://doi.org/10.1186/s12909-025-07592-y

Xu, J., Li, D., Lu, D., Chen, M., Chen, T., Sun, H., Qing, Y., Wu, C., & Wang, Y. (2026). Enhancing knowledge integration and self-directed learning in undergraduate medical education through an AI-based multimodal platform. *Advances in Medical Education and Practice, 17*, 605052. https://doi.org/10.2147/AMEP.S605052

Xu, L., Xu, Q., Liu, C., Chen, B., & Wang, C. (2025). Virtual standardized patients for improving clinical thinking ability training in residents: Randomized controlled trial. *JMIR Medical Education, 11*, e73196. https://doi.org/10.2196/73196

Yamamoto, A., Koda, M., Ogawa, H., Miyoshi, T., Maeda, Y., Otsuka, F., & Ino, H. (2024). Enhancing medical interview skills through AI-simulated patient interactions: Nonrandomized controlled trial. *JMIR Medical Education, 10*, e58753. https://doi.org/10.2196/58753

Yanagita, Y., Shikino, K., Yokokawa, D., Ishizuka, K., Tsukamoto, T., Li, Y., Uehara, T., & Ikusaka, M. (2025). Appropriate semantic qualifiers increase diagnostic accuracy when using a clinical decision support system: A randomized controlled trial. *BMC Medical Education, 25*(1), 711. https://doi.org/10.1186/s12909-025-07294-5

Yang, F., Wang, J., & He, L. (2026). Application of AI-assisted PBL-CBL combined teaching in

the diagnosis and treatment of post-herpetic neuralgia: A randomised controlled trial. *BMC Medical Education, 26*(1), 1149. https://doi.org/10.1186/s12909-026-09489-w

Yu, C., Li, F., Zhang, N., Hu, H., Huang, H., Wang, J., Tao, Y., & Wu, Y. (2026). Effectiveness of artificial intelligence–assisted peer teaching in orthopedic clinical education: Historical cohort study. *JMIR Medical Education, 12*, e87959. https://doi.org/10.2196/87959

Esta obra está bajo una licencia de Creative Commons Reconocimiento-No Comercial 4.0 Internacional. Copyright © Lissette Gardenia Hoyos Sanmartín y Eugenio Alejandro Ortega García.

Declaraciones éticas y editoriales del artículo
Contribución de los autores (Taxonomía CRediT). Lissette Gardenia Hoyos Sanmartín: conceptualización de la investigación, diseño metodológico, desarrollo del proceso investigativo, análisis formal de los datos, redacción del borrador original del manuscrito, revisión crítica del contenido científico y supervisión general del estudio. Eugenio Alejandro Ortega García: conceptualización de la investigación, diseño metodológico, desarrollo del proceso investigativo, análisis formal de los datos, redacción del borrador original del manuscrito, revisión crítica del contenido científico y supervisión general del estudio.
Declaración de conflicto de intereses Los autores declaran que no existe conflicto de intereses en relación con la investigación presentada, la autoría del manuscrito ni la publicación del presente artículo.
Declaración de financiamiento La presente investigación no recibió financiamiento específico de agencias públicas, comerciales o de organizaciones sin fines de lucro. En caso de existir financiamiento institucional o externo, este deberá ser declarado explícitamente por los autores en esta sección.
Declaración del editor El editor responsable certifica que el proceso editorial del presente artículo se desarrolló conforme a los principios de integridad científica, transparencia y buenas prácticas editoriales. El manuscrito fue sometido a un proceso de evaluación mediante revisión por pares doble ciego, garantizando la confidencialidad de la identidad de los autores y revisores durante todo el proceso de dictamen académico. Asimismo, el editor declara que el artículo cumple con los criterios científicos, metodológicos y éticos establecidos por la revista.
Declaración de los revisores Los revisores externos que participaron en la evaluación del presente manuscrito declaran haber realizado el proceso de revisión de manera objetiva, independiente y confidencial. Asimismo, manifiestan que no mantienen conflictos de interés con los autores ni con la investigación evaluada, y que sus observaciones y recomendaciones se fundamentan exclusivamente en criterios científicos, metodológicos y académicos.
Declaración ética de la investigación Los autores declaran que la investigación se desarrolló respetando los principios éticos de la investigación científica, garantizando la confidencialidad de los datos y el respeto a los participantes del estudio. En los casos en que la investigación involucre seres humanos, los procedimientos deben ajustarse a los principios éticos establecidos en la Declaración de Helsinki y a las normativas institucionales correspondientes.
Declaración sobre el uso de inteligencia artificial Los autores declaran que el uso de herramientas de inteligencia artificial, en caso de haberse utilizado durante el proceso de investigación o redacción del manuscrito, se realizó únicamente como apoyo técnico para mejorar la claridad del lenguaje o el análisis de información, manteniendo siempre la responsabilidad intelectual sobre el contenido del artículo. Las herramientas de inteligencia artificial no fueron utilizadas como autoras del manuscrito ni sustituyen la responsabilidad académica de los investigadores.
Disponibilidad de datos Los datos que respaldan los resultados de esta investigación estarán disponibles previa solicitud razonable al autor de correspondencia, respetando las normas éticas y de confidencialidad establecidas por la investigación.