
Inteligencia Artificial en Pediatría: avances y aplicaciones

Sección a cargo de Julián Battolla^a, María del Pilar Arias López^b y Gustavo Frei^a

De la pregunta clínica a la evidencia científica: un recorrido por las herramientas basadas en inteligencia artificial para apoyar la toma de decisiones clínicas

From clinical question to scientific evidence: an overview of artificial intelligence-based tools to support clinical decision-making

Introducción

La irrupción de los Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés) ha modificado rápidamente la forma en que muchos profesionales de la salud buscan y sintetizan la información médica. Diversas encuestas muestran un elevado grado de conocimiento e interés por estas herramientas entre los médicos: una encuesta realizada a 1 049 médicos de 50 países halló que el 86,5% refería un conocimiento de básico a avanzado de distintas herramientas basadas en Inteligencia Artificial (IA) y el 80,2% esperaba que mejorara su práctica clínica. Sin embargo, en el mismo estudio solo el 27,8% de los médicos había incorporado IA en la práctica, y el análisis atribuyó esa brecha, sobre todo, a factores estructurales (entrenamiento formal, inversión institucional) más que a actitudes individuales o variables demográficas.¹ Esta heterogeneidad se repite, con variaciones considerables, en otras publicaciones. Una encuesta realizada entre cirujanos generales de Alemania, Austria y Suiza informó que el 86,7% de los encuestados usaban IA en forma ocasional o regular, siendo ChatGPT® la herramienta dominante entre quienes usaban chatbots. Sin embargo, los propios autores advierten que la tasa de respuesta fue muy baja (8,7%; 323/3 831 encuestados) y la muestra desbalanceada, con probable sobrerrepresentación de los cirujanos previamente interesados en IA, lo que podría sobreestimar tanto el uso reportado como el optimismo hacia la herramienta.² Por el contrario, en una red hospitalaria de Nueva York, el 89,7% de los médicos refería conocer a los chatbots basados en LLM, pero solo el 21% los había incorporado efectivamente a su práctica, con la limitación de que este estudio no informaba la tasa de respuesta.³

En la misma línea, una encuesta realizada en Brasil mostró una marcada brecha entre adopción institucional e individual; solo el 4% de los establecimientos de salud reportó usar IA de forma institucional, mientras que el 16–17 % de médicos y enfermeras ya la usaban

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

individualmente por su cuenta, un patrón de adopción informal que precede a la implementación institucional consistente con lo observado a escala global.⁴

A nivel organizacional, la adopción de IA en el sector salud ha crecido de forma sostenida en los últimos 3 años, pero todavía permanece por detrás de otros sectores de la economía como finanzas, educación o servicios profesionales.⁵ Esta demora se documenta también en el ámbito público: entre los 50 países que respondieron una encuesta de la Oficina Regional para Europa de la Organización Mundial de la Salud, solo el 24% ofrecía formación en IA durante el servicio para su personal de salud.⁶

En este escenario ha surgido una aparente paradoja. La mayor parte de la literatura científica reciente compara plataformas clínicas especializadas, como OpenEvidence®, UpToDate Expert AI® o Dyna AI®, con modelos de propósito general como ChatGPT®. Sin embargo, los escasos datos disponibles sugieren que la adopción de modelos de lenguaje de propósito general por parte de los médicos precede ampliamente a la implementación de plataformas clínicas especializadas, cuya utilización permanece concentrada en determinados sistemas de salud e instituciones, aunque con excepciones notables donde la disponibilidad gratuita y verificada de una herramienta especializada invierte ese patrón, como lo demuestra un estudio de la Universidad de Emory donde el 51,9% de los profesionales encuestados utilizaba de forma autónoma OpenEvidence® frente a solo un 7,4% que recurría a ChatGPT®.⁷

Esta discrepancia entre las evaluaciones comparativas estandarizadas publicadas (*benchmarks*) y las modalidades reales de uso condiciona la interpretación de la evidencia disponible y constituye el punto de partida del presente artículo. Esta distinción no es menor para una revista de un hospital pediátrico: mientras que la literatura disponible se concentra desproporcionadamente en las herramientas especializadas (con financiamiento y fines comerciales), el uso más extendido y menos regulado de los chatbots de propósito general en la práctica real cuenta con escasa evidencia científica sobre su verdadero desempeño.

El objetivo de este artículo es ofrecer un panorama descriptivo y equilibrado de ambos tipos de herramientas (especializadas y de propósito real), sus luces y sus sombras, y proporcionar un marco para que el pediatra pueda posicionarse frente a ellas con criterio propio, sin sobreestimar ni descartar ninguna de las dos opciones.

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

Panorama descriptivo: ¿Qué herramientas existen y para qué se usan?

Conviene distinguir al menos cinco categorías de herramientas que suelen mencionarse indistintamente bajo el paraguas de "IA médica para la obtención de evidencia para la toma de decisiones clínicas" pero que difieren sustancialmente en su lógica de funcionamiento, su modelo de acceso, su disponibilidad real fuera de Estados Unidos y una importante heterogeneidad de funcionalidades y usuarios (Tabla 1).⁸

Las primeras cuatro categorías se utilizan, en mayor o menor medida, en el punto de atención; la quinta (asistentes de síntesis documental como Gemini Notebook[®]) tiene un uso más cercano al trabajo de investigación y editorial que a la asistencia del paciente. Estas últimas herramientas son cada vez más relevantes para autores y revisores, ya que permiten organizar y consultar un corpus propio de artículos, sin depender de una base de evidencia externa curada por un tercero.

En la tabla 1 se describen las herramientas disponibles y sus principales características. Las características más relevantes para nuestro contexto son la disponibilidad en español y el uso real. Un LLM de propósito general puede consultarse en español, sin costo ni verificación profesional alguna, desde cualquier país.

Entre las herramientas de soporte de decisión clínica (CDS) especializadas, conviene hacer una precisión verificada directamente por los autores al momento de escribir este artículo: OpenEvidence[®], en particular, ya admite el registro con matrícula profesional argentina y permite formular las consultas en español, una funcionalidad que probablemente no era así en versiones anteriores de la plataforma.

Aun así, su base documental subyacente (literatura indexada, guías, acuerdos de licencia con revistas) sigue estando construida principalmente sobre fuentes y contextos de práctica estadounidenses, lo que deja abierta una pregunta distinta a la de mero acceso: si el contenido recuperado y sintetizado resulta igualmente aplicable a un paciente pediátrico atendido bajo guías y recursos locales.

Esta distinción entre poder acceder a una herramienta y que la evidencia que esa herramienta recupera sea la relevante para nuestro sistema de salud, es la que consideramos más importante para la lectura crítica de esta sección, más que la barrera de acceso en sí misma.

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

Tabla 1. Categorías de herramientas basadas en inteligencia artificial utilizadas en la práctica clínica para obtención de evidencia

Categoría / ejemplos	Acceso y costo	Base de evidencia	Disponibilidad en español	Uso real estimado entre médicos
LLM de propósito general (ChatGPT®, Claude®, Gemini®)	Gratuitos en su versión básica; sin necesidad de registro profesional; acceso web y app sin restricción geográfica.	Conocimiento general incorporado en el entrenamiento. Sin curaduría médica dedicada ni citas verificadas por defecto.	Alta. Funcionan de forma nativa en español.	Mayoritario: es la forma más común en que un médico interactúa hoy con IAG, generalmente de manera informal.
Sistemas de Decisión Clínica basados en RAG entrenados con literatura biomédica (OpenEvidence®, UpToDate Expert AI®, DynaMed/Dyna AI®, ClinicalKey AI®)	OpenEvidence®: gratuito; requiere verificación profesional (NPI en Estados Unidos; matrícula profesional de otros países, incluida Argentina); ¹⁰⁻¹² UpToDate/DynaMed: pago, licencia institucional. ^{13,14}	OpenEvidence®: RAG entrenado en literatura indexada (PubMed, revistas con acuerdos de licencia), sin síntesis editorial previa; UpToDate®/DynaMed®/ClinicalKey AI®: RAG entrenado en contenido editorialmente curado y sintetizado por equipos médicos (revisiones temáticas, guías). ¹³⁻¹⁵	OpenEvidence® permite realizar consultas en español. La base documental subyacente, sin embargo, sigue siendo mayormente anglosajona.	Minoritario fuera de Estados Unidos, pero en expansión; alto crecimiento reciente en usuarios estadounidenses con NPI.
Herramientas de razonamiento diagnóstico (Glass Health®) ¹⁶	Modelo freemium; servicios básicos de forma gratuita y funciones avanzadas con costo. Generación de diagnósticos diferenciales a partir de la descripción de un caso.	Entrenado sobre guías clínicas curadas para estructurar diagnósticos diferenciales, no para responder consultas generales.	Limitada. Producto centrado en el mercado angloparlante.	Nicho; uso concentrado en clínicos que buscan estructurar razonamiento diagnóstico.
Sistemas agénticos / híbridos (ChatRWD®) ⁹	Acceso institucional/empresarial; no es una herramienta de consumo masivo.	Genera evidencia de mundo real bajo demanda mediante un motor de inferencia causal sobre datos clínicos reales.	Muy limitada, orientada a instituciones de Estados Unidos.	Muy minoritario; uso experimental o institucional, no de práctica clínica cotidiana.
Asistentes de síntesis documental (Gemini Notebook® y similares)	Gratuito o modelo freemium; sin registro profesional; el usuario carga sus propios documentos como fuentes (publicaciones, guías, historias clínicas anonimizadas).	No tiene base de evidencia propia: sintetiza únicamente el corpus de documentos que el usuario le proporciona, con citas trazables a esos documentos.	Alta, funciona de forma nativa en español sobre documentos redactados en cualquier idioma.	Creciente en tareas de investigación y trabajo editorial (revisión de literatura, preparación de manuscritos), no en la cabecera del paciente.

LLM: grandes modelos del lenguaje; IAG: inteligencia artificial generativa; RAG (Generación aumentada por Recuperación); NPI: National Provider Identifier (número de 10 dígitos, identificador único usado en el sistema de salud de Estados Unidos). Tabla construida por los autores, con datos respaldados por las referencias 9-16.

- Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG
- Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

Además, entre las herramientas mencionadas en la tabla 1 existen además diferencias comerciales relevantes para un lector clínico. OpenEvidence®, fundada en 2022 por el Dr. Daniel Nadler y Zachary Ziegler y lanzada públicamente en 2023, es gratuita para profesionales de la salud verificados y sostiene su modelo de negocio mediante acuerdos de licencia de contenido con editoriales como NEJM, JAMA y The Lancet.¹¹ El éxito de este modelo se refleja en su rápida expansión financiera y de mercado: respaldada por un nivel de adopción que ya abarca a cerca del 40% de los médicos en los Estados Unidos.¹² UpToDate Expert AI® (Wolters Kluwer)¹³ y DynaMed/Dyna AI® (EBSCO)¹⁴, en cambio, son productos con licenciamiento individual o institucional pago, con una trayectoria más larga como referencia clínica. Mientras que DynaMed fue fundado en 1995,¹⁴ UpToDate ha sido considerado durante mucho tiempo el estándar de referencia del sector, antes de incorporar capas de IA conversacional sobre su contenido preexistente.^{13,14} ClinicalKey AI®, de Elsevier, se apoya en el gran corpus de revistas científicas de la editorial y en tratados médicos de referencia mundial con licencia verificada, destacando de forma relevante para nuestra especialidad el Tratado de Pediatría de Nelson, junto a otras obras de referencia como la Medicina de Goldman-Cecil y la Cardiología de Braunwald.¹⁵ Glass Health®, por su parte, no compite en la recuperación de literatura general, sino que se especializa en generar diagnósticos diferenciales estructurados a partir de la descripción de un caso.¹⁶ Estas diferencias de modelo de negocio y de propósito no determinan por sí solas cuál de las herramientas rinde mejor pero sí son relevantes al momento de decidir cuál, si alguna, conviene incorporar institucionalmente.

Cómo funcionan los sistemas de decisión clínica basados en generación aumentada por recuperación: una explicación necesaria

Dado que buena parte de este artículo discute a OpenEvidence®, UpToDate Expert AI®, DynaMed/Dyna AI® y ClinicalKey AI®, y que no puede darse por sentado que todo lector conozca en detalle su funcionamiento interno, vale la pena detenerse en cuál es la diferencia real con un chatbot de propósito general, más allá del nombre comercial y del marketing de "IA especializada en medicina". Estas herramientas se basan en una técnica llamada **generación aumentada por recuperación** (RAG, por sus siglas en inglés, retrieval augmented generation).¹⁷

En términos simples, en lugar de que el modelo de lenguaje responda únicamente a partir de lo que "aprendió" durante su entrenamiento, como hace un LLM de propósito gene-

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

ral al ser consultado, el sistema primero realiza una búsqueda en una base documental acotada que, según la herramienta, puede ser literatura biomédica indexada sin síntesis editorial previa (el caso de OpenEvidence®, que recupera directamente de PubMed y revistas con acuerdos de licencia, como por ejemplo New England Journal of Medicine o Journal of the American Medicine Association) o un corpus de revisiones y guías curadas y sintetizadas editorialmente por equipos médicos (como UpToDate®, DynaMed® y ClinicalKey AI®) y recién después usa un LLM para redactar una respuesta que resuma y cite específicamente los documentos recuperados en esa búsqueda.¹⁷ La respuesta, en teoría, queda "anclada" a fuentes verificables, con enlaces o referencias que el usuario puede y debe verificar (Imagen 1).

Imagen 1. Funcionamiento de Grandes Modelos del Lenguaje vs. Sistemas de Soporte de Decisión Clínica para búsqueda y síntesis de evidencia científica



Fuente: elaboración propia

Esta arquitectura explica tanto su principal promesa como limitación. La promesa es que, al no generar la respuesta libremente sino recuperarla de una base documental acotada (curada editorialmente o simplemente indexada, según la herramienta), el riesgo de "alucinaciones" o citas inexistentes debería ser menor que en un LLM de propósito general. En efecto, varios estudios encuentran tasas de "alucinación" de citas más bajas para OpenEvidence® que para los chatbots generales.^{9,10} El límite es que la calidad de la respuesta dependerá de la interpretación de la pregunta, de la calidad, actualización y cobertura de la base documental consultada y del razonamiento sobre lo recuperado. Si la base consultada no incluye guías pediátricas locales, estudios en español, o literatura relevante para el contexto latinoamericano, el sistema simplemente no puede recuperarlos, por más sofisticado que sea

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

el modelo de lenguaje que redacta la respuesta final. En otras palabras: un RAG no es más "inteligente" que un LLM genérico, sino que es más "bibliotecario"; su valor depende de cuán buena y pertinente sea la biblioteca a la que tiene acceso.

Luces: ¿Qué funciona razonablemente bien?

La evidencia disponible, con las salvedades metodológicas que se mencionan a lo largo de esta sección, sugiere que tanto las herramientas especializadas como las de propósito general pueden aportar valor en tareas acotadas y bien definidas. Los LLM de propósito general muestran un desempeño creciente en conocimiento médico general y en tareas de comunicación (redacción, simplificación de lenguaje, generación de material educativo), mientras que las herramientas de RAG especializadas muestran ventajas en tareas de recuperación de literatura con citas verificables, sobre todo cuando la consulta es acotada, factual y del tipo que se resuelve típicamente en la cabecera del paciente (dosis, interacciones, guías de manejo específicas). Esta fortaleza comunicacional de los LLM generales está documentada empíricamente en un estudio ampliamente citado: al comparar respuestas de médicos y de ChatGPT® a preguntas de pacientes publicadas en un foro público de Reddit, los profesionales de la salud prefirieron las respuestas del chatbot en el 78,6% de los casos, calificándolas como más claras y empáticas.¹⁸ Sin embargo, hay que considerar que la comparación se dio en un foro de redes sociales, con respuestas médicas típicamente breves, y no en una consulta clínica real, lo que limita cuánto puede extrapolarse a la práctica asistencial. Hurt y cols. evaluando a OpenEvidence® en casos reales de atención primaria, encontraron alta calificación en claridad y respaldo en la evidencia, aunque con un patrón que se repite en buena parte de esta literatura: la herramienta refuerza la decisión que el médico ya había tomado, más que modificarla.^{9,10,19}

Sombras: ¿Dónde fallan las herramientas y por qué la propia evidencia es contro-versial?

La limitación más citada de los LLM de propósito general, la "alucinación" de citas o datos inexistentes, aparece de forma consistente en estudios de especialidades muy distintas entre sí, como ortopedia, radiología, hasta un estudio que incluyó 33 médicos de 17 especialidades y documentó explícitamente la generación de citas inexactas ante pedidos de fuentes, además de errores ocasionales pero categóricamente incorrectos en preguntas clínicas

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

abiertas, lo que sugiere que es una limitación estructural del tipo de arquitectura, no un defecto de un modelo en particular.²⁰⁻²² Pero el problema menos discutido, y quizás más relevante para esta sección, es otro: la propia evidencia acerca de cuál herramienta es "mejor" resulta sorprendentemente inestable de un estudio a otro, incluso cuando se comparan exactamente las mismas herramientas.

El ejemplo más claro es el contrapunto, publicado con pocas semanas de diferencia en 2026, entre Vishwanath y cols.²³ y Feng y cols.²⁴ El primero, un equipo independiente de NYU Langone evaluó a OpenEvidence® y UpToDate Expert AI® frente a Claude Opus 4.6, GPT-5.2 y Gemini 3.1 Pro en tres tareas: respuestas a preguntas médicas (MedQA), ítems que evaluaban concordancia entre las respuestas y opinión de clínicos expertos (HealthBench) y 100 preguntas realizadas por médicos a los LLM durante consultas clínicas reales extraídas de un sistema de ChatGPT® institucional. La investigación encontró que los tres modelos de propósito general superaron a ambas herramientas clínicas especializadas en las tres tareas.²³

El estudio publicado por Feng y cols., con participación de OpenEvidence® en el diseño y la recolección de datos (pero análisis estadístico predefinido y ciego), evaluó a la misma herramienta frente a versiones más recientes de los mismos tres modelos, usando 620 consultas reales efectivamente enviadas por médicos a la plataforma OpenEvidence® y 149 evaluadores especialistas seleccionados por tema, y encontró el resultado exactamente inverso: OpenEvidence® superó a los tres modelos generales en las cinco dimensiones evaluadas.²⁴

No presentamos este contrapunto para dirimir cuál de los dos estudios "tiene razón" sino como el ejemplo más ilustrativo de un problema metodológico de fondo: el resultado de cualquier evaluación comparativa de IA clínica depende fuertemente de decisiones de diseño que rara vez se hacen explícitas. De hecho, OpenEvidence® solicitó formalmente a Nature Medicine la retractación del primero, alegando contaminación de los benchmarks MedQA y HealthBench en el entrenamiento de los modelos de vanguardia; la revista respondió que "da la bienvenida al debate científico" y remitió la disputa a su proceso formal de "Matters Arising", que continuaba abierto al momento de escribir esta sección.²⁵

Al menos cuatro limitaciones son recurrentes en la literatura revisada:

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

- **El origen de las preguntas:** si provienen de usuarios que ya eligieron una de las herramientas evaluadas, el sesgo de selección favorece sistemáticamente a esa herramienta cuando se la evalúa "en su propio terreno".^{24,25}
- **El número y la especialización de los evaluadores:** paneles pequeños y de médicos generalistas (12 clínicos, en un caso) producen estimaciones distintas de paneles de expertos con mayor tamaño muestral, asignados por especialidad (149 médicos, en el otro).^{24,25}
- **El uso de humanos o modelos de lenguaje como jueces:** el propio estudio de Feng y cols. mostró que los LLM como jueces difieren sistemáticamente de los expertos humanos en el detalle de las clasificaciones, aunque ambos suelen coincidir en identificar el mejor modelo, un hallazgo con implicancias directas para HealthBench, que se apoya en jueces automatizados.²⁴
- **El formato de entrega de las preguntas:** James, Caronia y Savargaonkar mostraron, con preguntas de residencia pediátrica del estilo de las del examen de licencia médica de Estados Unidos (USMLE), que administrar las preguntas todas juntas o en lotes de diez alteraba la precisión medida de un rango de 38–90% (muy inconsistente) a 83–88% (convergente) para los mismos seis modelos, incluido OpenEvidence.²⁶

Ninguno de estos cuatro factores es una "trampa" deliberada; son simplemente grados de libertad que el campo todavía no ha estandarizado. Un ejemplo adicional de este patrón, en un tercer tipo de tarea, proviene del estudio de Low y cols.⁹ que informa que, al responder 50 preguntas clínicas reales, los LLM generales presentaron un rendimiento pobre frente a OpenEvidence® y un sistema agéntico con motor de inferencia causal, ChatRWD®.

Los LLM generales produjeron respuestas clínicamente relevantes y basadas en la evidencia en solo 2-10% de los casos 24 % vs. 58% para OpenEvidence® y CharRWD® respectivamente. Al contrario, Mauffrey y cols., evaluaron la generación de material educativo en ortopedia pediátrica y encontraron que ChatGPT-4® produjo textos más legibles y precisos que OpenEvidence®, aunque ninguno alcanzó el nivel de OrthoInfo, el estándar de referencia de la American Academy of Orthopaedic Surgeons.¹⁹ La Tabla 2 resume el diseño, los hallazgos principales y las limitaciones de cada uno de estos estudios mencionados.

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

Tabla 2. Síntesis de los estudios de evidencia comparativa revisados

Estudio	Diseño y Herramientas comparadas	Hallazgo principal	Limitaciones / Conflictos de Interés
Vishwanath y cols. ²³	3 tareas: 500 MedQA + 500 HealthBench + 100 RCQ; 12 clínicos evaluadores a ciegas. GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6 vs. OpenEvidence®, UpToDate Expert AI®	Los 3 LLM de propósito general superaron a ambas herramientas clínicas especializadas en las 3 evaluaciones; OpenEvidence®/UpToDate rindieron similar a Google AI Overview.	RCQ extraídas de un ChatGPT® institucional, no de usuarios de OpenEvidence®; solo 12 evaluadores generalistas.
Feng y cols. ²⁴	620 consultas reales de OpenEvidence® (Real-POCQi) + 187 de HealthBench; 149 médicos, 36 estados Claude Opus 4.8, Gemini 3.1 Pro, GPT-5.5 vs. OpenEvidence®	OpenEvidence® superó a los 3 LLM en las 5 dimensiones evaluadas (25–39 puntos porcentuales de diferencia); resultado robusto en análisis de sensibilidad.	Preguntas provienen de usuarios que ya eligieron OpenEvidence® (posible sesgo de selección); OpenEvidence® co-diseño y pagó la recolección de datos, aunque el análisis estadístico fue independiente.
Low y cols. ⁹	50 preguntas clínicas reales; 9 médicos evaluadores ChatGPT-4®, Claude 3 Opus, Gemini 1.5 Pro vs. OpenEvidence® vs. ChatRWD (agéntico)	Los LLM generales rara vez dieron respuestas relevantes y basadas en evidencia (2–10%); OpenEvidence 24%, ChatRWD 58%. Herramientas complementarias entre sí (66% combinadas).	Financiado por Atropos Health (desarrollador de ChatRWD); agradecimiento explícito a OpenEvidence® por soporte técnico.
Hurt y cols. ¹⁰	5 casos retrospectivos de atención primaria; 4 médicos evaluadores OpenEvidence®	Alta calificación en claridad, relevancia y respaldo en la evidencia; impacto bajo en la decisión clínica reforzó, pero no modificó los planes médicos.	Muestra muy pequeña; diseño retrospectivo sin grupo control prospectivo.
Mauffrey y cols. ¹⁹	51 condiciones ortopédicas pediátricas comunes; 4 cirujanos evaluadores (modalidad ciega) ChatGPT-4® vs. OpenEvidence® vs. OrthoInfo®	ChatGPT® logró materiales más legibles y precisos/completos que OpenEvidence®. OrthoInfo® siguió siendo el más preciso, pero con la peor legibilidad.	Un solo prompt por condición, sin evaluar consistencia; evaluadores (modalidad no ciega) Textos de IA más concisos).
Huang y cols. ²¹	106 preguntas del examen ACR DXIT 2022; comparación pareada 5 LLM multimodales (GPT-5.1, Gemini 3 Pro, Claude Sonnet 4.5, Grok 4.1, Bing Copilot GPT-5) vs. OpenEvidence® y DeepSeek® (solo texto)	En preguntas de solo texto, OpenEvidence® y DeepSeek® sin diferencia significativa entre sí. Todos los modelos multimodales cayeron ~35 puntos porcentuales en preguntas con imagen.	El artículo describe erróneamente a OpenEvidence® como desarrollado por una universidad sueca, dato fáctico incorrecto.

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

James cols. ²⁶	y	Estudio transversal; 77 MQAs de nivel de residencia pediátrica validados por panel de 3 pediatras (modalidad ciega). Comparación de 2 formatos de entrega: todo a la vez vs. lotes de 10. ChatGPT®, DeepSeek® AI, Gemini, Microsoft Copilot, Perplexity AI (all-at-once) + OpenEvidence (solo en lotes).	El formato de entrega en lotes de 10 redujo drásticamente la dispersión y estabilizó el rendimiento (convergencia a 83–88% de precisión) independientemente de la arquitectura del LLM.	Un solo set de preguntas pediátricas; evaluación en un único punto temporal; no incluyó grupo de control humano. Sin conflictos de interés declarados.
---------------------------	---	---	--	---

MedQA: preguntas y respuestas que evalúan conocimiento médico; HealthBench: items que miden concordancia entre las respuestas de los modelos y los expertos médicos; RCQ: preguntas clínicas reales (del inglés; Real Clinical Queries). Tabla construida por los autores. Se presentan en fuente negrita las herramientas evaluadas.

Dimensiones que los estudios de comparación de desempeño no capturan

Más allá de qué herramienta "gana" en cada estudio, hay dimensiones prácticas que casi ninguno de los estudios revisados mide, y que son precisamente las más relevantes para la toma de decisiones de un pediatra en nuestro medio:

- ✓ la latencia y el costo de cada herramienta en el flujo de trabajo real;
- ✓ la calidad y verificabilidad efectiva de las citas (no solo su presencia);
- ✓ la cobertura de guías de especialidad o recursos locales: Vishwanath y cols. reconocen explícitamente en su discusión que las evaluaciones comparativas (*benchmarks*) generales fallan en capturar la utilidad clínica en tareas profundamente sub especializadas y en flujos de trabajo e instituciones de salud específicas;
- ✓ la calidad real (no solo la existencia) de las respuestas en español, más allá de que el acceso y la consulta en ese idioma ya sean posibles en algunas de estas plataformas;
- ✓ la integración con los sistemas de historia clínica y las guías de las instituciones de salud latinoamericanas, hoy prácticamente inexistente, y consistente con las disparidades institucionales e infraestructurales ya documentadas en la región.⁴

Implicancias para la práctica pediátrica

De la evidencia reunida se desprenden tres implicancias prácticas razonables en la actualidad:

1. Cualquiera de las dos categorías de herramientas (chatbot de propósito general o CDS especializado) debe entenderse como apoyo a la decisión y no como sustituto de

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

ella, teniendo en cuenta el patrón repetido en la literatura que evidencia que estas herramientas tienden a reforzar antes que modificar el criterio clínico ya formado.

2. La evaluación crítica de cada respuesta generada, incluyendo la verificación efectiva de las citas cuando las hay, sigue siendo responsabilidad del médico, sin importar cuán "especializada" se presente la herramienta.
3. Ninguna de las herramientas revisadas en este artículo fue validada específicamente en población pediátrica latinoamericana, en español, ni en el contexto de nuestros sistemas de salud. La escasa evidencia pediátrica directa proviene de ortopedia estadounidense, por lo que cualquier extrapolación de estos hallazgos a nuestra práctica cotidiana debe hacerse con cautela explícita.¹⁹

Un dato de adopción, aunque no específico de IA generativa, ayuda a dimensionar la brecha entre lo que una evaluación comparativa (*benchmark*) mide y lo que efectivamente ocurre en la práctica: Exmann y cols. en una encuesta a 151 profesionales de siete países que atienden adultos mayores con condiciones crónicas complejas, encontraron que solo el 51,7 % usaba algún sistema de SDC y apenas 22,5% usaba herramientas predictivas, con variaciones enormes entre países que no se explicaban por la actitud individual hacia la tecnología sino, aparentemente, por factores estructurales del sistema de salud.²⁷ Es razonable esperar que la adopción real de IA generativa en pediatría latinoamericana enfrente barreras estructurales similares o mayores.

Consideraciones regulatorias y éticas

Un informe del Emergency Care Research Institute (ECRI), organismo independiente de seguridad en salud, nombró al mal uso de los chatbots de IA de propósito general como el riesgo tecnológico número uno para 2026, señalando específicamente a ChatGPT®, Claude®, Copilot®, Gemini® y Grok. 3.0®. Resulta fundamental aclarar que esta alerta se refiere particularmente a las herramientas de consumo general que los profesionales usan de manera informal. Por el contrario, las herramientas de decisión clínica especializadas (como OpenEvidence®) se diseñan bajo la lógica de fuentes de evidencia médica tradicionales. No obstante, ninguna de las dos categorías cuenta todavía con una validación independiente robusta y consistente entre tareas y poblaciones, como muestra el propio contrapunto Vishwanath-Feng reseñado más arriba.^{23,24} En este escenario, la actualización de enero de 2026 de las Recommendations del International Committee of Medical Journal Editors (ICMJE)

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

incorporó por primera vez una sección dedicada exclusivamente a la declaración del uso de IA, que exige que los autores divulguen activamente el uso de IA generativa en la redacción de manuscritos, tanto en la carta de presentación como en el propio texto, y que verifiquen rigurosamente cualquier contenido asistido por IA, dado que esta "puede generar texto de tono autorizado que sea incorrecto, incompleto o sesgado".²⁹

Conclusión

La pregunta relevante ya no es si los médicos utilizarán IA, sino cómo aprenderemos a utilizarla críticamente. La evidencia disponible sugiere que ninguna herramienta es consistentemente superior en todas las tareas y que los resultados de los estudios dependen tanto del diseño metodológico como del modelo evaluado. Para el pediatra, el desafío no consiste en elegir entre ChatGPT® u OpenEvidence®, sino en desarrollar las habilidades necesarias para interpretar, verificar e integrar la información generada por cualquiera de estas herramientas dentro del razonamiento clínico. En ese sentido, la alfabetización en IA probablemente se convierta en una competencia profesional tan importante como la lectura crítica de la literatura biomédica.

Bibliografía

1. Bold B, Serin O, Tantri Adhiatma L, et al. Global physician perspectives on artificial intelligence in healthcare across 50 countries and territories. *npj Digit Med*. 2026
2. Lünse S, Wisotzky EL, Renz-Kiefel L, et al. Artificial intelligence and chatbots in general surgery: a survey among surgeons in Germany, Austria and Switzerland. *Surg Endosc*. 2026 Jul 1. doi:10.1007/s00464-026-13044-5.
3. Solmonovich RL, Kouba I, Lee JY, et al. Physician awareness of, interest in, and current use of artificial intelligence large language model-based virtual assistants. *PLoS One*. 2025; 20(5):e0320749. doi:10.1371/journal.pone.0320749.
4. Portilho L, Marin HF. Artificial Intelligence: The Use and Application by Brazilian Healthcare Professionals. *Stud Health Technol Inform*. 2026; 336:705-709. doi:10.3233/SHTI260262.
5. Nguyen TD, Whaley CM, Simon K, et al. Adoption of Artificial Intelligence in the Health Care Sector. *JAMA Health Forum*. 2025; 6(11):e255029. doi:10.1001/jamahealthforum.2025.5029.
6. Adib K, Dunning HE, Letchford N, et al. Artificial intelligence in health systems of the WHO European region: implementation, applications, opportunities, and barriers. *Int J Med Inform*. 2026; 218:106532. doi:10.1016/j.ijmedinf.2026.106532.
7. Bagla P, Hanna J, Marthambadi B, et al. Patterns of AI Use in Clinical Work by Hospitalists: Survey Study. *J Med Internet Res*. 2026; 28:e85973. doi:10.2196/85973.
8. Laymouna M, Ma Y, Lessard D, et al. Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. *J Med Internet Res*. 2024; 26:e56930. doi:10.2196/56930.
9. Low YS, Jackson ML, Hyde RJ, et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digit health*. 2025;11:20552076251348850. doi:10.1177/20552076251348850.

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

10. Hurt RT, Stephenson CR, Gilman EA, et al. The Use of an Artificial Intelligence Platform Open-Evidence to Augment Clinical Decision-Making for Primary Care Physicians. *J Prim Care Community Health*. 2025;16:21501319251332215. doi:10.1177/21501319251332215.
11. Philip S, Kurian R. OpenEvidence. *J Med Libr Assoc*. 2026;114(1):86-87. doi:10.5195/jmla.2026.2247.
12. Carpenter Wellington PLLC. OpenEvidence closes \$250 million Series D funding round at \$12 billion valuation. Lexology [Internet]. 18 de febrero de 2026. [citado 24 Ago 2026]. Disponible en: [lexology.com/library/detail.aspx?g=e51cc1c8-f651-4ede-9659-d968cb54a7ab](https://www.lexology.com/library/detail.aspx?g=e51cc1c8-f651-4ede-9659-d968cb54a7ab).
13. Wolters Kluwer. Wolters Kluwer's new UpToDate Expert AI provides clinicians and health systems with the fast, reliable GenAI clinical decision support they need [comunicado de prensa en Internet]. 24 sept 2025. [citado el 24 de agosto de 2026]. Disponible en: <https://www.wolterskluwer.com/en/news/uptodate-expert-ai-genai-clinical-decision-support>
14. EBSCO. DynaMed vs. UpToDate: AI Clinical Decision Support Tools Comparison. Health-Notes blog [Internet]. 22 Dic 2025 [citado 24 ago 2026]. Disponible en: <https://about.ebsco.com/blogs/health-notes/dynamed-vs-uptodate-ai-clinical-decision-support-tools-comparison>
15. Elsevier. Elsevier Health launches ClinicalKey AI, an advanced clinical decision support tool for clinicians [comunicado de prensa en Internet]. 2024 feb 29 [24 agosto 2026]. Disponible en: <https://www.elsevier.com/about/press-releases/elsevier-health-launches-clinicalkey-ai-the-most-advanced-gen-ai-powered>
16. Glass Health. Clinical AI Resources for Clinicians [Internet]. 2026 [citado 24 ago 2026]. Disponible en: <https://glass.health/resources>
17. Amazon Web Services. ¿Qué es RAG (generación aumentada por recuperación)? [Internet]. 2026 [citado 24 ago 2026]. Disponible en: <https://aws.amazon.com/what-is/retrieval-augmented-generation/>
18. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838.
19. Mauffrey O, Lashani S, Heyer JH, et al. The ABCs of PEMs: Using Artificial Intelligence to Enhance the Readability of Patient Educational Materials in Pediatric Orthopaedics. *J Pediatr Orthop Soc N Am*. 2025; 13:100273. doi:10.1016/j.jposna.2025.100273.
20. Walters C, Young P. Artificial Intelligence in Orthopedic Trauma: A Narrative Review. *Cureus*. 2026; 18(4):e107574. doi:10.7759/cureus.107574.
21. Huang KA, Samvelian D, Xu AL, et al. Comparative Performance of Seven Mainstream Large Language Models on the 2022 American College of Radiology Diagnostic Imaging In-Training Examination. *Cureus*. 2026; 18(4):e106486. doi:10.7759/cureus.106486.
22. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw Open*. 2023; 6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.
23. Vishwanath K, Alyakin A, Ghosh M, et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat Med*. 2026;32(7):2405-2409. doi:10.1038/s41591-026-04431-5.
24. Feng J, Patel V, Heagerty P, et al. Expert Evaluation of Clinical AI Tools on Real Point-of-Care Clinical Queries. *arXiv:2606.28960v1 [cs.AI]*. 2026 Jun 27. [citado 24 ago 2026]. doi: <https://doi.org/10.48550/arXiv.2606.28960>

a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG

b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG

25. Giles B. ChatGPT, Gemini, Claude outperform clinical AI tools: Study. Becker's Hospital Review. [Internet]16 Junio 2026. [citado 24 ago 2026]. Disponible en: beckershospitalreview.com/healthcare-information-technology/ai/chatgpt-gemini-claude-beat-clinical-ai-tools-study/
26. James V, Caronia C, Savargaonkar R. Input structure-driven instability and convergence in large language model clinical reasoning: a formative study using pediatric residency-level MCQs. Sci Rep. 2026;16(1):17829. doi:10.1038/s41598-026-48326-4.
27. Exmann CJC, Hiltunen AM, Haavisto I, et al. Uptake of Clinical Decision Support Systems Among Health Care Professionals in Six European Countries and the United States: Cross-Sectional Survey Within the I-CARE4OLD Project. J Med Internet Res. 2026; 28:e85071. doi:10.2196/85071.
28. Abdelwahed AE, Abd El-Nasser M, Heih OQ, et al. Public attitudes and practices toward using AI chatbots for healthcare assistance: a multinational cross-sectional study. BMC Health Serv Res. 2026; 26:335. doi:10.1186/s12913-025-13832-0
29. International Committee of Medical Journal Editors (ICMJE). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated January 2026, Section V (Artificial Intelligence). Disponible en: [icmje.org/recommendations/](https://www.icmje.org/recommendations/).
30. ECRI. Top 10 Health Technology Hazards for 2026: Executive Brief. [Internet] 21 de enero de 2026. [citado 24 ago 2026]. Disponible en: <https://home.ecri.org/blogs/ecri-thought-leadership-resources/top-10-health-technology-hazards-for-2026-executive-brief>

Forma de citar: Battolla J, Arias López MP, Frei G. De la pregunta clínica a la evidencia científica: un recorrido por las herramientas basadas en inteligencia artificial para apoyar la toma de decisiones clínicas. Rev. Hosp. Niños (B. Aires) 2026; 68 (302):288-302

Declaración sobre uso de IA en este manuscrito: de acuerdo con las guías GAMER (Generative Artificial Intelligence tools in MEDical Research), los autores declaramos el uso de herramientas basadas en IA generativa (IAG) en el proceso de preparación de este manuscrito. Se empleó el asistente de síntesis documental Gemini Notebook® (basado en los modelos de lenguaje de Google, versión de agosto de 2026), ChatGPT® (OpenAI, versión GPT-4o-2024-05-13, a través de su interfaz de usuario web) y Claude® (Anthropic, versión Claude 3.5 Sonnet, a través de su interfaz de usuario web) durante el período comprendido entre el 1 de julio y el 25 de agosto de 2026. Se implementaron técnicas de ingeniería de prompts iterativas y refinamiento de consultas de forma manual para la extracción de metadatos de los estudios, estructuración de las comparaciones y verificación cruzada de la coherencia fáctica de la literatura médica de soporte. Las herramientas se utilizaron como soporte metodológico para la búsqueda bibliográfica, la síntesis descriptiva preliminar de los estudios comparativos y la revisión estilística del texto original en español redactado por los autores. La contribución de la IA generativa se localiza específicamente en la estructuración del manuscrito, la construcción de la Tabla 2 (benchmarks) y la revisión de estilo general. No se desarrollaron ni ajustaron modelos de IA propios para este estudio. Los autores contrastaron manualmente cada dato, porcentaje, hallazgo y cita bibliográfica generada o sugerida por las herramientas de IAG contra los documentos médicos primarios indexados en la biblioteca de origen, corrigiendo de forma activa las discrepancias fácticas. Los autores asumen la responsabilidad total y final por la exactitud, integridad y originalidad de todo el contenido publicado. No se introdujo información confidencial de pacientes ni datos de historias clínicas.

- a. Médico pediatra. Unidad Consultorios Externos matutinos, Consultorios de Mediana y Alta Complejidad, HNRG
- b. Médica pediatra. Unidad de Cuidados Intensivos Pediátricos, HNRG