22º congreso actualización pediatría 2026

Viernes 6 de marzo de 2026
Seminario:
Aplicación práctica de la IA
en la consulta de Pediatría
de Atención Primaria

Moderadora:
Irene García de Diego
Pediatra. CS Aranjuez. Madrid. Vocal del Comité de Docencia de la Asociación Madrileña de Pediatría de Atención Primaria (AMPap).

Ponente/monitor:
- **Ernesto Barrera Linares**
  Pediatra. CS San Blas. Parla. Madrid.

Textos disponibles en
www.aepap.org

¿Cómo citar este artículo?

Barrera Linares E. La inteligencia artificial como asistente del pediatra de Atención Primaria: evidencias, aplicaciones y riesgos. En: AEPap (ed.). Congreso de Actualización en Pediatría 2026. Madrid: Lúa Ediciones 3.0; 2026. p. xxx-xxx.

La inteligencia artificial como asistente del pediatra de Atención Primaria: evidencias, aplicaciones y riesgos

Ernesto Barrera Linares
Pediatra. CS San Blas. Parla. Madrid
ernestobarreral@gmail.com

RESUMEN

Este seminario evalúa la aplicación de la Inteligencia Artificial (IA) en el ámbito de la salud, basándose en una síntesis de evidencia científica reciente. El objetivo es proporcionar al Pediatra de Atención Primaria (AP) una valoración crítica de las oportunidades y riesgos de los grandes modelos de lenguaje (LLM) y el Deep Learning. La evidencia muestra un rendimiento a nivel de experto en la interpretación de imágenes médicas (ej., detección de neumonía[1]) y en la automatización de tareas administrativas, logrando alta precisión en la extracción de datos[2,3] y reducción del tiempo de documentación[4].

No obstante, el rendimiento en el razonamiento clínico es inconsistente. Existen riesgos de fiabilidad, como las "alucinaciones" (fabricación de referencias[5]) y sesgos algorítmicos[6,7]. El cuerpo de evidencia es predominantemente de calidad metodológica moderada a baja, con una notable falta de validación prospectiva[8] y datos pediátricos específicos[9,10]. Se presenta una clasificación de herramientas IA según niveles de evidencia[30] y un método estructurado (RECORD[24]) para optimizar la interacción. La confianza es alta para la automatización, pero baja a moderada para la toma de decisiones clínicas de alto riesgo. La recomendación es adoptar un enfoque de "aumento, no reemplazo", utilizando la IA como un asistente inteligente bajo la supervisión y el juicio crítico del profesional.

INTRODUCCIÓN

La irrupción de la Inteligencia Artificial (IA) está transformando rápidamente el panorama médico. Para el Pediatra de Atención Primaria (AP), comprender el alcance global de esta tecnología es fundamental, ya que los avances validados en otras especialidades (como la radiología) son a menudo precursores de las herramientas que llegarán a nuestra práctica.

Además, muchos modelos, especialmente los grandes modelos de lenguaje (LLM), son de propósito general. Esto significa que no han sido específicamente entrenados o validados en poblaciones pediátricas[9,10]. Por lo tanto, conocer su rendimiento y sus limitaciones en el ámbito general de la salud es indispensable para realizar una valoración crítica antes de considerar su uso con nuestros pacientes.

La síntesis de evidencia reciente proporciona una visión integral sobre la IA, desde el diagnóstico hasta la gestión administrativa y los desafíos éticos.

EVIDENCIAS DE APLICACIÓN CLÍNICA

La IA ha demostrado una capacidad robusta en tareas específicas de clasificación y automatización.

Soporte al diagnóstico y decisión

Los modelos de *Deep Learning* muestran un rendimiento a nivel de experto o superior en la interpretación de imágenes médicas. Por ejemplo, encontramos estudios que reportan altas precisiones en la detección de neumonía en radiografías de tórax[1], en la clasificación de mamografías[11], o incluso superando a radiólogos en la diferenciación de lesiones pulmonares[12]. Esta alta precisión también se observa en señales biológicas, como en la detección de convulsiones neonatales (EEG)[13] o en la clasificación de arritmias (ECG)[14].

En el ámbito del razonamiento clínico complejo, los LLM muestran resultados divergentes. Algunos estudios son prometedores, sugiriendo que modelos como ChatGPT-4o pueden igualar o superar a los médicos en escenarios pediátricos complejos[15] o mostrar alta precisión en enfermedades raras[16]. Sin embargo, este rendimiento no es universal. Cuando se evalúan los LLM de propósito general con el rigor científico que aplicamos a cualquier otra prueba diagnóstica, encontramos limitaciones significativas. Por ejemplo, un estudio que evaluó la capacidad de los LLM para la valoración fotográfica de la escoliosis concluyó que, por el momento, no son suficientemente aceptables, reportando una especificidad del 0% y errores sistemáticos que excedían ampliamente la tolerancia clínica[17].

Automatización de tareas (la aplicación más segura)

La automatización de la documentación es, hoy por hoy, una de las aplicaciones más maduras y de mayor impacto. Los LLM pueden extraer información clínica de texto no estructurado con muy alta precisión[2,3], por ejemplo, identificando fiebre en lactantes[18] o hallazgos en informes de TC[19]. En la práctica, esto se traduce en una reducción de la carga administrativa. La IA puede generar borradores de informes de radiología reduciendo el tiempo de lectura en un 42%[4], y los resúmenes médicos que genera son comparables en completitud a los escritos por médicos, pero se obtienen en segundos[20].

Educación médica y del paciente

La IA también se perfila como una herramienta formativa. Se utiliza para generar preguntas de examen (MCQ) de alta calidad[21] y para crear simulaciones con pacientes virtuales que mejoran las habilidades de comunicación[22]. Para nuestros pacientes, los LLM pueden generar borradores de materiales educativos[23] y se está explorando su uso como *chatbots* de apoyo (ej. en salud mental juvenil).

RIESGOS CRÍTICOS, SESGOS Y FIABILIDAD

La implementación clínica segura se ve comprometida por importantes barreras.

Alucinaciones y fiabilidad

Un problema grave es la "alucinación" o fabricación de información. Al evaluar la precisión de las referencias

bibliográficas, se han encontrado tasas de fabricación de hasta el 60.6% en algunos modelos[5]. Además, la fiabilidad del conocimiento es inconsistente; un estudio documentó cómo un LLM no fue capaz de actualizar su conocimiento sobre las guías clínicas de timpanostomía a lo largo de un año, repitiendo los mismos errores[25].

Sesgos algorítmicos

La IA puede perpetuar los sesgos humanos presentes en los datos. Se han identificado sesgos raciales, donde los modelos proponen tratamientos de menor calidad para minorías raciales[6], y sesgos de género y etnia en la generación de imágenes[7].

CARTOGRAFÍA DE LA IA EN SALUD: ESTRATEGIA Y HERRAMIENTAS

Navegar hoy por el ecosistema de la Inteligencia Artificial requiere un mapa conceptual. Para ordenar este entorno tan cambiante, se propone un enfoque basado en tres pilares: un marco de clasificación (la Pirámide de Evidencia 5.0), una selección de recursos y una técnica de interacción (Método RECORD).

Antes de situar las herramientas, es vital distinguir los conceptos técnicos que definen su fiabilidad:

- IA fundacional: el "cerebro" puro (ej. ChatGPT, Claude, Gemini). Razona y escribe muy bien, pero su conocimiento está "congelado" en su fecha de entrenamiento, además de que son modelos de lenguaje grande (LLM) entrenados con fuentes de información generales, no específicas del dominio de conocimiento sanitario.

- RAG (Generación Aumentada por Recuperación): conecta ese cerebro a una fuente externa para buscar datos antes de responder. Puede ser una búsqueda en internet o en tus propios archivos.

- *Whitelist* (lista blanca): un tipo de RAG seguro que solo busca en fuentes preaprobadas (guías clínicas, revistas Q1), ignorando el ruido de la web general.

- Agentes: sistemas autónomos que planifican pasos complejos (buscar, leer, comparar y resumir) sin intervención humana constante.

Marco conceptual: la Pirámide de Evidencia 5.0

Partiendo del modelo "EBHC Pyramid 5.0" de Alper y Haynes[29], proponemos una adaptación que integra las nuevas herramientas de IA según el grado de procesamiento de la información y su aplicabilidad clínica directa.

A. La base: búsqueda y síntesis (investigación). Aquí la IA actúa como un "agente de investigación" que reduce horas de lectura a minutos de análisis.

- Búsqueda semántica: herramientas como Undermind o Scite.ai superan la coincidencia de palabras clave. Entienden conceptos complejos y permiten rastrear la literatura evaluando la calidad metodológica (ej. detectando si un estudio ha sido retractado).

- Agentes de síntesis: plataformas como Elicit o Consensus son capaces de leer miles de *abstracts*. Su valor reside en generar matrices de evidencia o tablas comparativas automáticas, ideales para responder preguntas de contexto o *background* (¿qué se sabe sobre X?) y visualizar el estado del arte.

B. La cima: sumarios y sistemas (*Point-of-Care*). En la consulta directa, necesitamos respuestas inmediatas, verificables y seguras.

- RAG clínico y académico: herramientas como Open Evidence funcionan con "listas blancas", respondiendo solo si encuentran la respuesta en fuentes biomédicas de alta calidad. En esta categoría destaca también Perplexity (en su modo Pro/Academic), que permite realizar búsquedas profundas en la web académica citando las fuentes en tiempo real.

- RAG personalizado (gestión del conocimiento): herramientas como Google NotebookLM permiten "chatear" con sus propios PDF (protocolos del centro, guías descargadas). Además de asegurar que la respuesta se basa exclusivamente en los

documentos fiables seleccionados (sin riesgo de invención externa), permiten generar formatos alternativos (como audio-resúmenes), apoyando de forma sorprendente un trabajo cognitivo de síntesis que hasta la fecha no existía.

- Inteligencia ambiental: sistemas de escucha activa que, con el debido consentimiento, filtran la conversación clínica, eliminan la charla social y estructuran automáticamente la nota en la historia clínica, devolviendo al médico el contacto visual con el paciente.

(Nota: para una visualización interactiva de estas categorías y su evolución, se está desarrollando el proyecto "Pirámide de Evidencia IA", disponible como recurso complementario en la web del autor).

Usos transversales: personalización y modelos fundacionales

Más allá de la búsqueda de evidencia, los Modelos Fundacionales (como ChatGPT, Claude Sonnet o Gemini) son potentes aliados para tareas de redacción, resumen, traducción y adaptación de textos (ej. simplificar un informe para los padres).

Una de las grandes revoluciones recientes es la capacidad de personalización mediante los GPT (en OpenAI) o las Gems (en Gemini). Estas funciones permiten al pediatra crear "mini-asistentes" preconfigurados para tareas repetitivas sin saber programar.

- Ejemplo: se puede crear un "GPT revisor de analíticas" instruido para que, cada vez que le peguemos unos resultados (estrictamente anonimizados), genere automáticamente una versión explicativa en lenguaje sencillo y empático para la familia, pistas para la interpretación clínica, o sugerencias para ampliar el estudio.

El lenguaje: ingeniería de *prompts* y método RECORD

Para interactuar con estos modelos fundacionales y evitar respuestas vagas o "alucinaciones", debemos ser precisos. Una IA generativa debe tratarse como un residente muy capaz pero propenso a la fabulación si no se le dan instrucciones claras.

El futurista Alvin Toffler escribió: "La pregunta correcta es generalmente más importante que la respuesta correcta a la pregunta equivocada". En medicina, la calidad del output depende de la calidad del *prompt*. Para ello, se propone el marco mnemotécnico RECORD (Barrera-Linares, 2024):

- R - Rol: define quién debe ser la IA. (Ej.: "Actúa como pediatra experto en AP...").
- E - Escenario: sitúa el contexto. (Ej.: "...en una consulta saturada...").
- C - Contexto del caso: aporta los datos. (Ej.: "Niño de 4 años, fiebre de 3 días...").
- O - Objetivo: define la tarea exacta. (Ej.: "Dame 3 diagnósticos diferenciales.").
- R - Restricciones: pon límites. (Ej.: "No inventes datos.").
- D - Diseño: especifica el formato. (Ej.: "Una tabla comparativa.").

La estrategia: el Modelo Sándwich

La adopción de estas herramientas no implica sustitución, sino colaboración. El flujo de trabajo más seguro es el Modelo Sándwich:

1. Pan superior (humano): el pediatra define la estrategia y elige la herramienta (¿Necesito un dato de Open Evidence o redactar con un GPT personalizado?).
2. Relleno (IA): la máquina realiza el procesamiento masivo.
3. Pan inferior (humano): la verificación es innegociable (el principio *human-in-the-loop* que la literatura considera esencial en IA clínica). Debemos

tratar el output de la IA como una prueba diagnóstica: conociendo su sensibilidad y especificidad.

Este enfoque híbrido nos permite aprovechar la "inteligencia extendida" sin renunciar a la responsabilidad ética y legal que siempre será humana... por el momento.

EVALUACIÓN CRÍTICA GLOBAL DE LA EVIDENCIA

El cuerpo de evidencia sobre IA en medicina es predominantemente de calidad metodológica moderada a baja, con un fuerte predominio de estudios de validación retrospectiva. Existen sesgos sistemáticos que limitan la generalización, destacando el sesgo geográfico/poblacional (la mayoría de los estudios provienen de instituciones de altos ingresos) y una extrema heterogeneidad metodológica que dificulta las comparaciones[8].

CONCLUSIONES Y RECOMENDACIONES

La evidencia demuestra que la IA ha superado la fase conceptual. Sin embargo, su implementación debe ser cautelosa y guiada por el juicio clínico. En lugar de una tabla de certeza formal, podemos resumir el estado actual de la siguiente manera:

- La confianza es alta para el uso de la IA en la automatización de tareas administrativas (ej. resumir historias, extraer datos), donde el ahorro de tiempo es significativo y el riesgo clínico es bajo[4,19,20].
- La evidencia es moderada para su uso como apoyo en la interpretación de imágenes médicas y señales biológicas (ej. radiografías, ECGs), donde muestra alta precisión, pero aún requiere validación prospectiva[1,11-14].
- La confianza es baja a moderada para el uso de LLM en el razonamiento clínico complejo o el diagnóstico de novo, debido a su inconsistencia[15-17].
- Los motores con RAG como Perplexity ofrecen confianza moderada-alta al proporcionar fuentes verificables.
- Los modelos fundacionales con ventanas de contexto ampliadas (hasta 1M tokens) muestran potencial para síntesis de documentación extensa[26], aunque requieren validación específica en pediatría.
- La evidencia es alta en que los modelos actuales presentan sesgos y limitaciones de fiabilidad (alucinaciones, conocimiento no actualizado) que impiden su uso autónomo en decisiones de alto riesgo[5-7,25].

La recomendación pragmática y segura es adoptar un enfoque de "aumento, no reemplazo". El profesional siempre es el responsable final y debe "evaluar críticamente cada sugerencia, contrastarla con guías de práctica clínica y aplicar su propio juicio y experiencia".

Existen Gaps críticos que limitan la implementación segura de la IA en pediatría:

1. Faltan ensayos clínicos prospectivos y aleatorizados que evalúen el impacto de la IA en los resultados de los pacientes y los costos sanitarios[8,27].
2. No existe un consenso sobre marcos de evaluación estandarizados que incluyan equidad, interpretabilidad y seguridad[28,29].
3. Las poblaciones minoritarias y pediátricas están subrepresentadas en los conjuntos de datos, lo que plantea serias preocupaciones sobre la equidad[9,10,30].

BIBLIOGRAFÍA

1. Mustapha B, Zhou Y, Shan C, Xiao Z. Enhanced Pneumonia Detection in Chest X-Rays Using Hybrid Convolutional and Vision Transformer Networks. Curr Med Imaging. 2025;21:e15734056326685.
2. Patel P, Davis C, Ralbovsky A, Tinoco D, Williams CYK, Slatter S, *et al.* Large Language Models Outperform Traditional Natural Language Processing Methods in Extracting PatientReported Outcomes in Inflammatory Bowel Disease. Gastro Hep Adv. 2025;4:100563.

3. Lindsay SE, Madison CJ, Ramsey D, Doung YC, Gundle KR. De Novo Natural Language Processing Algorithm Accurately Identifies Myxofibrosarcoma From Pathology Reports. Clin Orthop Relat Res. 2025;483:80-7.

4. Hong EK, Roh B, Park B, Jo JB, Bae W, Park JS, *et al*. Value of Using a Generative AI Model in Chest Radiography Reporting: A Reader Study. Radiology. 2025;314(3):e241646.

5. Güneş YC, Cesur T, Camur E. Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties. Diagn Interv Radiol. 2025.

6. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. NPJ Digit Med. 2025;8(1):332.

7. Currie G, Hewis J, Hawk E, Rohren E. Gender and Ethnicity Bias of Text-to-Image Generative Artificial Intelligence in Medical Imaging, Part 1: Preliminary Evaluation. J Nucl Med Technol. 2024; 52(4):356-9.

8. Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI Agents in Clinical Medicine: A Systematic Review. medRxiv. 2025.08.22.25334232.

9. Golder S, Connor KO, López G, Tatonetti NP, González G. Leveraging Unstructured Data in Electronic Health Records to Detect Adverse Events from Pediatric Drug Use: A Scoping Review. Annu Rev Biomed Data Sci. 2025;8:227-50.

10. Navarathna N, Kanhere A, Gomez C, Isaiah A. Artificial intelligence in pediatric otolaryngology: A state-of-the-art review of opportunities and pitfalls. Int J Pediatr Otorhinolaryngol. 2025;194:112369.

11. Al Mansour AGM, Alshomrani F, Alfahaid A, Almitairi ATM. MammoViT: A Custom Vision Transformer Architecture for Accurate BIRADS Classification in Mammogram Analysis. Diagnostics (Basel). 2025;15(3):285.

12. Zhang G, Shang l, Li S, Zhang J, Zhang Z, Zhang X, *et al*. Non-enhanced CT deep learning model for differentiating lung adenocarcinoma from tuberculoma: a multicenter diagnostic study. Eur Radiol. 2025;35(12):8116-25.

13. Tuncer T, Dogan S, Tasci I, Tasci B, Hajiyeva R, *et al*. TATPat based explainable EEG model for neonatal seizure detection. Sci Rep. 2024;14:26688.

14. Bi S, Lu R, Xu Q, Zhang P. Accurate Arrhythmia Classification with Multi-Branch, Multi-Head Attention Temporal Convolutional Networks. Sensors (Basel). 2024;24(24):8124.

15. Del Monte F, Barolo R, Circhetla M, Delmonaco AG, Castagno E, Pivetta E, *et al*. Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. Front Digit Health. 2025;7:1624786.

16. Zhong W, Liu Y, Liu Y, Yang K, Gao H, Yan H, *et al*. Performance of ChatGPT-4o and Four Open-Source Large Language Models in Generating Diagnoses Based on China's Rare Disease Catalog: Comparative Study. J Med Internet Res. 2025;27:e69929.

17. Aydin C, Duygu OB, Karakas AB, Er E, Gokman G, Ozturk AM, *et al*. Clinical Failure of General-Purpose AI in Photographic Scoliosis Assessment: A Diagnostic Accuracy Study. Medicina (Kaunas). 2025;61(8):1342.

18. Aronson PL, Kuppermann N, Mahajan P, Nielsen B, Olsen CS, Meeks HD, *et al*. Natural Language Processing to Identify Infants Aged 90 Days and Younger With Fevers Prior to Presentation. Hosp Pediatr. 2025;15(1):e1-e15.

19. Wihl J, Rosenkranz E, Schramm S, Berberrich C, Griessmair M, Woznicki P, *et al*. Data extraction from free-text stroke CT reports using GPT-4o and Llama-3.3-70B: the impact of annotation guidelines. Eur Radiol Exp. 2025;9(1):61.

20. Schoonbeek RC, Workum JD, Schuit SCE, Hoekman AH, Mehri T, Doornberg JN, *et al*. Quality and efficiency of integrating customised large language model-generated summaries versus physician-written summaries: a validation study. BMJ Open. 2025;15(9):e099301.

21. Chen M, Ma J, Cui X, Dai Q, Hu H, Wu Y, *et al*. Advancing medical education in cervical cancer control with large language models for multiple-choice question generation. Med Teach. 2025;12:1-11.

22. Chen PJ, Liou WK. ChatGPT-driven interactive virtual reality communication simulation in obstetric nursing: A mixed-methods study. Nurse Educ Pract. 2025;85:104383.

23. King RC, Samaan JS, Haquang J, Bharani V, Margolis S, Srinivasan N, *et al*. Improving the Readability of Institutional Heart Failure-Related Patient Education Materials Using GPT-4: Observational Study. JMIR Cardio. 2025;9:e68817.

24. Barrera-Linares E. (2025, agosto 13). Método RECORD para escribir Prompts en ChatGPT. Zenodo. https://doi.org/10.5281/zenodo.16813419

25. Durgut O, Dikici O. Does ChatGPT update itself? Accuracy of ChatGPT in tympanostomy tube guidance: A comparative analysis with current literatura. Eur Arch Otorhinolaryngol. 2025.

26. Chen D, Parsa R, Swason K, Nunez JJ, Critch A, Bitterman DS, *et al*. Large language models in oncology: a review. BMJ Oncol. 2025;4(1):e000759.

27. D'Antonoli TA, Bluethgen C, Cuocolo R, Klontzas ME, Ponsiglione A, Kocat B. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. Diagn Interv Radiol. 2025.

28. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, *et al*. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak. 2024;24(1):357.

29. Alper BS, Haynes RB. EBHC pyramid 5.0 for accessing preappraised evidence and guidance. Evid Based Med. 2016;21(4):123-5.

30. Barrera E. Pirámide de Evidencia 5.0 y Herramientas IA. 2025. [Fecha de acceso 15 dic 2025]. Disponible en https://ernestobarrera.github.io/piramide-evidencia-ia.html