Inteligencia artificial aplicada a búsqueda de evidencia en salud mental

José María Rosales-Crespo

Enfermero Especialista en Salud Mental, UGC Salud Mental, Hospital Universitario de Puerto Real

Doctor en Ciencias de la Salud, Universidad de Cádiz

Miembro de la Asociación Española de Enfermería en Salud Mental (AEESME)

En salud mental, la búsqueda de evidencia científica responde a la necesidad de tomar decisiones clínicas en un contexto de incertidumbre, heterogeneidad y alta carga asistencial. El volumen de literatura obliga a sostener estrategias rigurosas —términos, filtros, fuentes y registro del proceso— para aproximarse a la mejor evidencia disponible. En este escenario, la inteligencia artificial (IA) puede reducir carga operativa y acelerar fases del proceso de búsqueda, siempre que su uso se integre en un proceso documentado y bajo supervisión profesional[1].

El problema aparece cuando se confunde una herramienta cognitiva con un método científico. El riesgo no es usar *prompts*; es delegar juicio clínico y metodológico en una respuesta con apariencia de rigor. La Organización Mundial de la Salud (OMS) ha advertido que los modelos generativos pueden producir respuestas convincentes que no garantizan veracidad y subraya la necesidad de gobernanza, transparencia y uso apropiado en contextos sanitarios[2]. Además, se ha mostrado de forma empírica que los modelos de lenguaje pueden generar referencias inexactas o inexistentes, lo que obliga a verificar de manera sistemática antes de aceptar cualquier respuesta como evidencia[3].

En síntesis, la IA no mejora la evidencia por sí sola. Lo determinante sigue siendo el criterio clínico y metodológico del profesional; sin ese criterio, la IA no corrige errores, los amplifica.

Delimitar el alcance de la IA en la búsqueda de evidencia

La IA aplicada a la búsqueda de evidencia no es homogénea. Integra funciones distintas con perfiles de riesgo diferentes y tratarlas como equivalentes conduce a errores metodológicos. Por ello, para valorar su utilidad, conviene distinguir, al menos, tres usos.

- *Recuperación bibliográfica (information retrieval, IR).* La IA actúa como soporte operativo en la identificación y la organización de resultados: sugiere sinónimos y descriptores; ayuda a ampliar términos; facilita tareas organizativas como la *deduplicación*; y puede agrupar resultados por temas. Esto ahorra tiempo y fricción en búsquedas amplias. Pero no convierte

una búsqueda en rigurosa ni sustituye el juicio científico. Sin una estrategia predefinida y sin registro auditable, la IA solo acelera decisiones opacas y aumenta el sesgo de selección.

- *Automatización del cribado en revisiones.* En el cribado de títulos y resúmenes, algunas herramientas emplean enfoques de *active learning* para priorizar registros por probabilidad de relevancia. Esto reduce carga y tiempo. El riesgo aparece si se interpreta como sustituto de los criterios de inclusión/exclusión o se usa sin control de calidad. La priorización algorítmica no garantiza exhaustividad ni corrige una estrategia de búsqueda mal planteada; solo reorganiza el trabajo. Bien utilizada, puede ser muy útil, pero exige supervisión humana y trazabilidad del proceso. En la práctica, existen herramientas que priorizan registros, pero mantienen la decisión en manos del revisor y documentan el procedimiento; ASReview es un ejemplo.[4]
- *IA generativa (modelos de lenguaje, LLM).* En este uso, la IA genera texto en forma de resúmenes, respuestas y síntesis que simulan conclusiones. Es el escenario de mayor riesgo, ya que una respuesta puede ser coherente y convincente sin ser verificable ni reflejar la lógica del método científico. El problema no es la redacción, es aceptar como conclusión una respuesta no contrastada. Por eso, los marcos recientes sobre uso responsable de IA en síntesis de evidencia insisten en transparencia, justificación del uso, responsabilidad humana, y verificación de afirmaciones y referencias[1].

En consecuencia, su uso es defendible cuando acelera recuperación, organización y cribado sin desplazar decisiones metodológicas, y cuando el proceso queda documentado y es revisable.

Tres errores frecuentes cuando falta método

Hay tres errores recurrentes cuando falta método, y la IA los amplifica. El primero es creer que una mala pregunta se compensa con herramientas. Una estrategia incompleta o una delimitación clínica pobre no se arreglan formulando mejor la consulta. El riesgo no es solo perder estudios relevantes, sino construir conclusiones sobre un conjunto seleccionado de forma no controlada. El segundo es creer que la coherencia equivale a veracidad; en modelos generativos, la autoridad puede ser solo apariencia. Se han documentado referencias ficticias incluso con formato correcto[3]. El tercero es

pensar que un *prompt* bien escrito salva una búsqueda sin procedimiento. Puede ordenar la interacción y pedir fuentes. Pero no valida resultados ni sustituye un proceso reproducible con criterios explícitos y registro.

En salud mental, el problema central no es tecnológico, sino clínico. La IA puede organizar información y generar texto coherente, pero la relevancia y la aplicabilidad dependen del caso y del contexto. Que un resultado se parezca a la pregunta no garantiza su utilidad clínica: hay que valorar su calidad y, sobre todo, si es aplicable en la situación concreta. Un modelo generativo no puede resolver esa integración contextual; decidir qué es aplicable exige juicio profesional y conocimiento del contexto asistencial[2,5].

En este punto aparece una trampa frecuente: confundir forma con método. Un texto puede sonar técnico, tener estructura y concluir con seguridad sin que exista una búsqueda reproducible, criterios explícitos o verificación. Si no se distingue entre una respuesta bien formulada y un proceso científicamente defendible, la IA no aporta rigor; aporta falsa seguridad. En nuestro campo, donde la aplicabilidad está mediada por variables contextuales —historia de trauma, adherencia, apoyo social, condiciones de vida, continuidad asistencial—esta confusión es especialmente problemática.

Prompts como andamiaje cognitivo

Conviene precisar el papel real de los *prompts*. No son una técnica de investigación ni sustituyen al método; son una forma de externalizar el razonamiento y fijar límites al interactuar con un sistema que genera texto. Bien usados, ayudan a explicitar objetivos, criterios y exclusiones; a exigir que cite fuentes y declare incertidumbre; y a introducir controles (comprobaciones cruzadas, alternativas, límites de generalización) que facilitan la verificación posterior[1].

Por eso conviene separar lo que aportan y lo que no. Aportan claridad del objetivo, petición explícita de fuentes y límites, y señales de control, por ejemplo, exigir que toda afirmación esté vinculada a un documento localizable. Pero no sustituyen una estrategia reproducible de búsqueda ni protegen frente a alucinaciones si no se verifica. Como aplicación práctica, puede ser útil estructurar los *prompts* con una regla simple para fijar límites y facilitar la verificación.

- **Tarea:** Especificar qué necesitas (términos MeSH, sinónimos, sugerir criterios, redactar un borrador de estrategia de búsqueda).
- **Contexto:** Indicar para qué y para quién (población, dispositivo, objetivo).
- **Criterios y límites:** Definir qué debe/no debe hacer (no inventar datos o referencias, declarar incertidumbre, no extrapolar).
- **Fuentes y verificación:** Exigir respuestas verificables (fuentes localizables, diferenciar términos controlados/libres, y comprobar citas y afirmaciones).
- **Formato:** Pedir la estructura de la salida, la extensión y el nivel de detalle deseado.

Riesgos específicos en salud mental

En salud mental el margen de error es más estrecho: el significado clínico depende del contexto, del vínculo y de la interpretación del lenguaje.

Ambigüedad y narrativa. Los datos suelen llegar en forma de relato y significado personal. Dos descripciones similares pueden corresponder a problemas distintos según la historia, el entorno y el momento evolutivo. Los modelos de lenguaje tienden a homogeneizar patrones, lo que puede proporcionar respuestas plausibles pero mal ajustadas a la singularidad del caso. La literatura sobre aplicaciones de LLM en salud mental insiste en equilibrar oportunidades con evaluación y mitigación de riesgos precisamente por esta fragilidad contextual[5,6].

Autoridad algorítmica y relación terapéutica. Una respuesta bien escrita puede desplazar el juicio clínico y erosionar la deliberación compartida. En salud mental esto pesa más porque la intervención no es solo aquello que se hace, sino cómo se acompaña, cómo se sostiene la incertidumbre y cómo se protege la alianza terapéutica. La OMS advierte del riesgo de confianza indebida en respuestas convincentes y subraya gobernanza, transparencia y uso apropiado de modelos generativos en contextos sanitarios[2].

Sesgo y estigma. El lenguaje en salud mental es especialmente vulnerable a sesgos (moralización, estereotipos, etiquetado) y a daños por simplificación. Por eso, además de método y verificación,

se necesita vigilancia activa del lenguaje y de los supuestos implícitos, con controles explícitos y responsabilidad profesional[7].

Confidencialidad y datos. En la práctica clínica, el riesgo también se relaciona con la privacidad. No es aceptable introducir información identificable de pacientes en modelos o plataformas no autorizadas. Si se usa IA, debe hacerse con herramientas aprobadas por la organización con minimización y anonimización, y manteniendo el control profesional sobre qué se comparte y qué se decide, en línea con los principios de gobernanza y uso apropiado en salud[2].

Propuesta práctica: mínimos de seguridad metodológica

En salud mental, la clave es separar qué usos son defendibles y cuáles no. La Tabla 1 distingue, de forma práctica, qué tareas son defendibles cuando el proceso es auditable y cuáles concentran mayor riesgo si se delegan sin verificación.

Tabla 1. Uso recomendado de IA en búsqueda de evidencia: qué sí y qué no.

Uso recomendado si el proceso es auditable	Uso no recomendado si se delega sin verificación
Proponer sinónimos y descriptores, traducir términos y sugerir combinaciones.	Pedir "la mejor evidencia" sin estrategia reproducible ni fuentes declaradas.
Redactar un borrador de estrategia para ejecutar en bases reales y guardar registros.	Pedir "hazme la revisión" sin corpus, sin trazabilidad del cribado y sin verificación.
Priorizar el cribado con *active learning* manteniendo supervisión humana y documentación del procedimiento.	Generar bibliografía final sin comprobación: elevado riesgo de referencias ficticias o inexactas.
Resumir artículos ya seleccionados o extraer variables con una plantilla, contrastando con el texto completo.	Elaborar síntesis o conclusiones a partir de resúmenes generados sin comprobación.
Apoyar la redacción con revisión humana y declaración de uso.	Redactar recomendaciones clínicas finales sin evaluación crítica ni verificación.

Para que estos usos sean defendibles, conviene aplicar mínimos de calidad que mantengan el proceso reproducible. A modo de guía práctica, la siguiente lista resume esos requisitos operativos para mantener registro y verificación cuando se usa IA. Puede aplicarse antes de la búsqueda, durante el cribado y al redactar la síntesis final.

Lista de calidad mínima del proceso (7 ítems)

✓ Pregunta explícita formulada en PICO/PECOS o equivalente.

✓ Fuentes declaradas de forma explícita.

✓ Estrategia de búsqueda guardada y reproducible.

✓ Criterios de inclusión/exclusión predefinidos.

✓ Comprobación y verificación de citas.

✓ Declaración del uso de IA (herramienta/modelo, fase y verificación).

✓ Responsabilidad final humana (quién verifica y asume la decisión).

Como enfermero especialista en salud mental, la incorporación de la IA a la práctica basada en evidencia me interesa en un sentido estrictamente pragmático: ganar eficiencia sin perder rigor. Bien integrada, puede actuar como un apoyo útil en tareas operativas de búsqueda, organización y apoyo a la redacción, liberando tiempo para lo que no es automatizable en nuestro ámbito: el juicio clínico, la deliberación compartida y el cuidado de la relación terapéutica. Para que ese potencial sea clínicamente defendible, su uso debe quedar siempre sometido a comprobaciones explícitas, registro del proceso y responsabilidad profesional. Solo así la IA puede convertirse en una aliada real para los profesionales de salud mental, desde distintas disciplinas, que aspiran a sostener decisiones más informadas, transparentes y contextualizadas.

Bibliografía

1. Flemyng E, Noel-Storr A, Macura B, Gartlehner G, Thomas J, Meerpohl JJ, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence. Environ Evid. 2025; 14:20. doi: https://doi.org/10.1186/s13750-025-00374-5
2. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: World Health Organization; 2025.
3. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. J Med Internet Res. 2024; 5(26):e52935. doi: https://doi.org/10.2196/52935
4. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021; 3:125-133. doi: https://doi.org/doi:10.1038/s42256-020-00287-7.
5. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. npj Digit Med. 2025; 8:230. doi: https://doi.org/10.1038/s41746-025-01611-4.
6. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Eichstaedt JC, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. npj Ment Health Res. 2024; 3:12. doi: https://doi.org/10.1038/s44184-024-00056-z.
7. Apakama DU, Nguyen K-A-N, Hyppolite D, Soffer S, Mudrik A, Ling E, et al. Identifying Bias at Scale in Clinical Notes Using Large Language Models. Mayo Clin Proc Digit Health. 2025; 3(4):100296. doi: https://doi.org/10.1016/j.mcpdig.2025.100296.