PUBLICACIONES DE LA
ACADEMIA NACIONAL DE
MEDICINA DE MÉXICO

ACTUALIDADES EN INTELIGENCIA ARTIFICIAL

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Número 8

Sesgos humanos vs sesgos de la inteligencia artificial

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra

Médicos Internistas
Universidad de Guadalajara
Instituto Mexicano del Seguro Social (IMSS)
Colegiados y Certificados (CMMI)
Fundadores del TICC
15 de abril de 2026

Revisión estructurada crítica con enfoque para médicos de TICC en la Clínica PalencIA Formato IMRyD · síntesis narrativa basada en evidencia reciente y verificable Abril 2026.

Documento preparado para lectura docente y discusión clínica. La revisión prioriza literatura de 2021–2026, con énfasis en revisiones sistemáticas, metaanálisis, marcos metodológicos y ensayos aleatorizados aplicables a la práctica médica.

Mensaje central. En medicina, la IA no sustituye el sesgo humano; lo redistribuye. Puede mitigarlo en tareas concretas, pero también puede amplificarlo, volverlo silencioso y escalarlo a nivel sistémico si se implementa sin auditoría, validación local y supervisión clínica.

Resumen

La comparación entre sesgos humanos y sesgos de la inteligencia artificial (IA) ha dejado de ser un debate abstracto y se ha convertido en un problema operativo para el ejercicio clínico.

El juicio humano sigue siendo vulnerable a sesgos cognitivos clásicos —anclaje, disponibilidad, confirmación, cierre prematuro y exceso de confianza—, mientras que la IA puede introducir o amplificar sesgos derivados de los datos, del etiquetado, de la selección de variables, de la deriva temporal y del contexto de despliegue[7-10]. La evidencia reciente sugiere que la IA no elimina el sesgo médico; puede mitigarlo en escenarios específicos o empeorarlo cuando el modelo está sesgado, mal calibrado o mal integrado al flujo clínico[11-15]. En particular, ensayos contemporáneos muestran que una IA sesgada puede deteriorar de manera significativa la precisión diagnóstica del médico, incluso cuando ofrece explicaciones[11,13,14]. Por ello, la pregunta clínicamente relevante no es qué sistema está "más sesgado", sino cómo interactúan ambos y bajo qué condiciones esa interacción mejora o empeora la seguridad del paciente. Para TICC en la Clínica PalencIA, la postura más prudente es adoptar un modelo híbrido: clínicos entrenados en sesgos cognitivos, algoritmos auditados por equidad y desempeño, y vigilancia continua de automatización acrítica, deriva y daño diferencial[5,11,15].

Palabras clave: sesgos cognitivos, inteligencia artificial, sesgo algorítmico, razonamiento clínico, seguridad del paciente, medicina interna, apoyo a la decisión clínica.

Introducción

El razonamiento clínico humano y la IA comparten una característica incómoda: ambos pueden fallar de forma predecible. El clínico falla porque decide bajo presión de tiempo, información incompleta, fatiga, sobrecarga cognitiva y contexto organizacional. La IA falla porque aprende de datos históricos sesgados, etiquetas imperfectas, poblaciones poco representativas o entornos que cambian después del entrenamiento[7-10]. La diferencia práctica es que el error humano suele ser local y discontinuo, mientras que el error algorítmico puede ser reproducible, silencioso y escalable[9,10].

En medicina interna, urgencias y hospitalización, esta distinción es crítica. Un sesgo humano puede alterar un caso; un sesgo algorítmico puede afectar simultáneamente miles de decisiones si el sistema se despliega a gran escala. A ello se suma un tercer fenómeno: la interacción humano–IA. Cuando el clínico recibe una

recomendación algorítmica, puede ocurrir corrección, compensación o arrastre. El arrastre —automation bias— es particularmente peligroso porque convierte un error técnico en un error clínico compartido[11-14].

Para los médicos de TICC en la Clínica PalencIA, el tema no debe abordarse como una confrontación entre humano y máquina, sino como un problema de arquitectura de decisiones. La pregunta central es cómo diseñar sistemas que reduzcan el error total, respeten la autonomía profesional, mejoren la equidad y mantengan la responsabilidad clínica en manos del médico tratante[5,11,15].

Métodos

Se elaboró una revisión estructurada narrativa con formato IMRyD y orientación crítica. La estrategia bibliográfica se diseñó para PubMed utilizando términos MeSH y operadores booleanos, y se complementó con corroboración documental en registros indexados, páginas editoriales y resúmenes estructurados de artículos localizados mediante búsqueda secundaria. Se priorizaron revisiones sistemáticas, metaanálisis, guías o extensiones metodológicas internacionales y ensayos controlados aleatorizados publicados entre enero de 2021 y abril de 2026, en inglés y español[1-6].

Se excluyó literatura no verificable, artículos de revistas señaladas como depredadoras y referencias sin DOI corroborable. Cuando un marco metodológico relevante era anterior o ligeramente externo al periodo ideal, se mantuvo por su valor normativo para la interpretación del tema, como ocurre con PRISMA 2020, CONSORT-AI y SPIRIT-AI[1-3].

La búsqueda base en PubMed puede reproducirse con sintaxis como:

("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Clinical Decision Support Systems"[Mesh]) AND (bias OR fairness OR "algorithmic bias" OR explainability OR trust OR "automation bias")
("diagnostic errors"[Mesh] OR "clinical reasoning" OR "cognitive bias" OR "decision making") AND (physicians OR "internal medicine" OR mergency)
("large language models" OR "generative artificial intelligence") AND (medicine OR diagnosis) AND (bias OR reasoning OR safety)
("systematic review"[Publication Type] OR meta-analysis OR randomized) AND (medical AI OR cognitive bias OR algorithmic bias)

La síntesis se organizó en cuatro dominios: sesgos cognitivos humanos, sesgos algorítmicos, interacción humano–IA y marcos de gobernanza y reporte.

Resultados

1. Sesgos humanos: persistencia, patrones y límites del des-sesgo

La literatura reciente confirma que los sesgos cognitivos siguen siendo frecuentes en medicina clínica. Una revisión de alcance en medicina interna identificó 41 sesgos estudiados, con predominio de anclaje, disponibilidad, confirmación y cierre prematuro7. En paralelo, una revisión sistemática y metaanálisis sobre herramientas de razonamiento cognitivo mostró una mejoría modesta pero significativa de la precisión diagnóstica, lo que sugiere que el sesgo humano puede atenuarse mediante estrategias deliberativas, aunque no eliminarse[8].

En atención crítica prehospitalaria, otra revisión identificó 28 sesgos, destacando anclaje, framing, disponibilidad, confirmación, exceso de confianza, cierre prematuro y sesgo de omisión[9]. El mensaje clínico es consistente: el médico no decide en un vacío lógico, sino dentro de un entorno cargado de presión, expectativas y atajos mentales que son útiles para la eficiencia, pero peligrosos cuando el caso es ambiguo, raro o dinámico[7-9].

2. Sesgos de la IA: de la base de datos al daño diferencial

Los sesgos de la IA suelen originarse en etapas menos visibles: selección y representatividad de datos, sesgo de medición, errores de etiquetado, variables proxy, deriva temporal y falta de validación externa10,15. Una revisión sistemática sobre modelos basados en expedientes electrónicos subrayó precisamente la heterogeneidad de fuentes de sesgo y la inconsistencia en las estrategias de mitigación reportadas[10].

En áreas sensibles a inequidad, como enfermedad cardiovascular y acceso diferencial a servicios, las revisiones sistemáticas recientes muestran que muchos algoritmos reproducen o amplifican disparidades raciales y étnicas cuando aprenden de datos históricos o de proxies socioeconómicos en lugar de necesidad clínica real[15,16]. Esto obliga a entender el sesgo algorítmico no como un defecto puramente técnico, sino como un problema sociotécnico que integra diseño, gobernanza, medición y contexto de uso[10,15,16].

3. Explicabilidad, confianza y falsa seguridad

La explicabilidad no garantiza seguridad clínica. Una revisión sistemática mostró que las explicaciones algorítmicas pueden aumentar, disminuir o no modificar la confianza del clínico según la calidad de la explicación, su forma de presentación y el contexto de uso[17]. De manera concordante, una revisión reciente sobre confianza del personal sanitario en sistemas AI-CDSS identificó ocho dominios decisivos: transparencia, capacitación, usabilidad, fiabilidad, validación, ética, diseño centrado en el usuario y capacidad de control o personalización[18].

La lección operativa es clara: un sistema "explicable" pero mal validado puede seguir siendo peligroso. Explicar no sustituye a calibrar, auditar y monitorizar. De hecho, ciertas explicaciones pueden inducir una falsa sensación de objetividad y aumentar la aceptación de una salida incorrecta[17,18].

4. Interacción humano–IA: donde el riesgo realmente se materializa

Los estudios experimentales ofrecen el hallazgo más clínicamente relevante. En un ensayo aleatorizado publicado en JAMA, la exposición a predicciones de una IA estándar mejoró modestamente la precisión diagnóstica, pero la exposición a una IA sistemáticamente sesgada la redujo de forma importante. Además, añadir explicaciones no corrigió de manera significativa el deterioro inducido por el sesgo del modelo[11].

En mamografía, un estudio en Radiology mostró automation bias en lectores con distintos niveles de experiencia cuando recibían sugerencias BI-RADS de IA[12]. En un ensayo aleatorizado sobre razonamiento diagnóstico con modelos de lenguaje, el uso de un LLM no produjo una mejora robusta y consistente del desempeño del médico frente a recursos habituales[13]. Más recientemente, un ensayo en escenario de dolor torácico demostró que la asistencia algorítmica puede modificar la decisión clínica y reconfigurar el patrón de sesgo del médico[14].

En conjunto, estos estudios indican que una IA sesgada no solo puede equivocarse, sino arrastrar al clínico a equivocarse con ella. Esto convierte a la supervisión humana en condición necesaria, pero no suficiente: la supervisión debe ser competente, entrenada y protegida frente a la sobreconfianza en el sistema[11-14].

Tabla 1. Comparación operativa entre sesgos humanos y sesgos de la IA

Dimensión	Sesgos humanos	Sesgos de IA	Riesgo principal	Mitigación razonable
Origen	Heurísticos, presión asistencial, fatiga, contexto y emociones	Datos sesgados, etiquetas imperfectas, variables proxy y drift	Error diagnóstico o terapéutico	Forcing strategies + auditoría del modelo
Temporalidad	Episódico y variable entre clínicos/turnos	Reproducible, persistente y escalable	Daño sistemático en subgrupos	Monitoreo continuo y revalidación
Transparencia	A veces verbalizable retrospectivamente	Con frecuencia opaca o seudotransparente	Falsa objetividad	Explicabilidad útil, no cosmética
Dependencia del contexto	Muy alta	Alta pero subestimada por el usuario	Extrapolación insegura	Validación local y externa
Corrección del error	Puede mejorar con reflexión y segunda opinión	Requiere rediseño, recalibración o retiro	Persistencia del fallo	Gobernanza técnica y clínica
Impacto sobre equidad	Variable según clínico y entorno	Puede amplificar inequidades históricas	Sesgo contra subpoblaciones	Evaluación por subgrupos
Interacción mutua	Puede desconfiar o sobreconfiar en IA	Puede reforzar sesgos humanos devolviendo salidas plausibles	Automation bias	Uso como segundo lector, no árbitro final

Discusión

La comparación entre sesgo humano y sesgo de IA suele formularse como si se tratara de elegir al árbitro menos defectuoso. La evidencia reciente no respalda ese planteamiento. El problema real es la suma, compensación o potenciación mutua entre ambos sistemas de decisión. El sesgo humano es ecológico: depende del turno, del cansancio, del flujo de trabajo y del contexto emocional. El sesgo de IA es infraestructural: depende del dataset, del etiquetado, del umbral de decisión, de la población de entrenamiento y del modo de implementación[7-10].

Por eso, una IA con muy buen rendimiento promedio puede seguir siendo inaceptable clínicamente si falla de manera sistemática en grupos concretos, si genera confianza desproporcionada o si no resiste la validación externa. Del mismo modo, un clínico experto puede beneficiarse de la IA en tareas delimitadas —por ejemplo, segunda lectura, priorización o cribado—, pero empeorar su juicio si la integración al flujo asistencial favorece el cierre prematuro o la delegación acrítica[11,12,17,18].

Este punto tiene especial peso en medicina interna y urgencias. En pacientes complejos, polimedicados o con presentaciones atípicas, la plausibilidad estadística no siempre coincide con la verdad clínica. Un modelo puede acertar la mayoría de los casos comunes y, sin embargo, fallar justo donde el internista más necesita apoyo: enfermedades raras, fenotipos atípicos, multimorbilidad o contextos con datos incompletos. En

esos escenarios, el valor de la IA depende menos de su promedio y más de su comportamiento en el borde del error[10-15].

Tabla 2. Marcos metodológicos y de reporte útiles para proyectos de IA clínica y revisiones en TICC

Marco	Propósito	Aplicación práctica
PRISMA 2020	Reporte transparente de revisiones	Ordenar búsqueda, selección, síntesis y limitaciones de la revisión
AMSTAR-2 / GRADE	Valoración metodológica y certeza	Juzgar si una revisión secundaria o recomendación es realmente confiable
CONSORT-AI	Ensayos clínicos con IA	Exigir reporte de versión del algoritmo, interacción humano–IA y errores
SPIRIT-AI	Protocolos de ensayos con IA	Definir desde el diseño cómo se evaluará la intervención algorítmica
RAISE	Uso responsable de IA en síntesis de evidencia	Hacer explícita la supervisión humana, trazabilidad y riesgos
GAMER	Reporte del uso de IA generativa en investigación	Declarar con transparencia dónde y cómo se usó IA generativa

La implementación segura de IA clínica no puede separarse de estos marcos. PRISMA 2020 orienta la transparencia de revisiones; CONSORT-AI y SPIRIT-AI obligan a reportar elementos específicos de intervenciones algorítmicas; GAMER y RAISE aportan trazabilidad cuando se usa IA generativa o automatización en síntesis de evidencia1–6,19. Un sistema técnicamente brillante pero metodológicamente opaco es, en términos clínicos, un sistema no confiable[2,3,5,6,19].

Implicaciones prácticas para médicos de TICC en la Clínica PalencIA

- No use la IA como juez final; úsela como segundo lector, sintetizador o generador de hipótesis.
- Antes de aceptar una recomendación algorítmica, pregunte: ¿en qué población fue entrenada?, ¿se validó localmente?, ¿cómo falla?, ¿en qué subgrupos pierde desempeño?
- Cuando la sugerencia del sistema coincida demasiado rápido con su impresión inicial, active una pausa cognitiva: esa concordancia puede reforzar anclaje y cierre prematuro.
- Documente discordancias entre criterio clínico e IA. Esas discordancias son material de auditoría, docencia y mejora continua.
- Evalúe no solo exactitud promedio: mida calibración, sensibilidad por subgrupos, drift, tasa de sobreconfianza y eventos de automation bias.

Limitaciones e incertidumbres

Persisten varias incertidumbres. Primero, todavía son escasos los estudios que comparan de forma directa sesgo humano y sesgo de IA en condiciones clínicas reales. Segundo, muchos ensayos utilizan viñetas o escenarios simulados, lo que limita la extrapolación a pacientes complejos, multimórbidos o inestables. Tercero, la evidencia sobre equidad y desempeño en poblaciones latinoamericanas es insuficiente. Cuarto, la explicabilidad continúa siendo heterogénea y no toda explicación mejora la confianza calibrada del clínico[11-18].

Por estas razones, cualquier implementación en medicina interna, urgencias o docencia clínica debe asumirse inicialmente como una intervención de riesgo moderado-alto hasta demostrar seguridad, validez local y utilidad incremental frente al estándar de cuidado[10,15,17,18].

Conclusiones

Los sesgos humanos y los sesgos de la IA no compiten; se acoplan. El sesgo humano nace del razonamiento bajo presión. El sesgo de la IA nace de datos, diseño y gobernanza. La fortaleza del clínico es la contextualización; la fortaleza de la IA es la consistencia. La debilidad del clínico es la variabilidad; la debilidad de la IA es la escalabilidad del error[7-15].

La literatura reciente no respalda la idea de que la IA, por sí sola, corrija el sesgo médico. Sí respalda, en cambio, un modelo híbrido: clínicos entrenados en sesgos cognitivos, algoritmos auditados por equidad y desempeño, validación local, vigilancia continua y obligación explícita de disentir del modelo cuando el contexto clínico lo exija[8,10-18].

Para TICC en la Clínica PalencIA, la conclusión operativa es directa: la IA debe ampliar el juicio clínico, no reemplazarlo; y todo sistema que no pueda ser auditado, explicado de forma útil y corregido en tiempo real no debe ocupar un lugar decisorio central en la práctica médica[5,17,18].

Bibliografía

1. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.

2. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x.
3. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7.
4. Luo X, Moher D, Lin L, Ge L, Wang Y, Wang R, et al. Reporting guideline for the use of Generative Artificial Intelligence in medical research: the GAMER statement. BMJ Evid Based Med. 2025. doi:10.1136/bmjebm-2025-113825.
5. Flemyng E, Ge L, Burns J, Kaltenbach S, Lotfi T, Lee YY, et al. Position statement and recommendations on the use of artificial intelligence in evidence synthesis and related reporting guidelines. Environ Evid. 2025;14:20. doi:10.1186/s13750-025-00374-5.
6. Thomas J, Kaltenbach S, Ge L, Flemyng E, et al. Responsible AI in Evidence SynthEsis (RAISE): guidance and recommendations. OSF Preprints. 2025. doi:10.17605/OSF.IO/FWAUD.
7. Loncharich MF, Trowbridge RL, Hallen S. Cognitive biases in internal medicine: a scoping review. Diagnosis (Berl). 2023;10(3):205-214. doi:10.1515/dx-2022-0120.
8. Staal J, Cozijnsen MA, Collette EH, Harskamp RE, Heringhaus C, Stengel D, et al. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. BMJ Qual Saf. 2022;31(12):899-910. doi:10.1136/bmjqs-2022-014865.
9. Awanzo A, Thompson J. Cognitive biases in clinical decision-making in prehospital critical care; a scoping review. Scand J Trauma Resusc Emerg Med. 2025;33(1):101. doi:10.1186/s13049-025-01415-1.
10. Chen F, Hsu Y, Kirkendall ES, et al. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based clinical AI models. J Am Med Inform Assoc. 2024. doi:10.1093/jamia/ocae060.
11. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, Wiens J, Sjoding MW. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295.
12. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4):e222176. doi:10.1148/radiol.222176.
13. Goh E, Lee J, Ang S, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969.
14. Goh E, Bunning B, Khoong EC, Gallo RJ, Milstein A, Centola D, Chen JH. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Commun Med (Lond). 2025;5(1). doi:10.1038/s43856-025-00781-2.
15. Cau R, Vogl TJ, Kuan T, et al. Addressing hidden risks: systematic review of artificial intelligence biases across racial and ethnic groups in cardiovascular diseases. Eur J Radiol. 2025. doi:10.1016/j.ejrad.2024.111867.
16. Haider SA, Deb A, Taha D, et al. The Algorithmic Divide: a systematic review on AI-driven racial disparities in healthcare. J Racial Ethn Health Disparities. 2024. doi:10.1007/s40615-024-02237-0.
17. Rosenbacke R, Wewetzer C, Kassaee A, et al. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI. 2024;3:e53207. doi:10.2196/53207.
18. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res. 2025;27:e69678. doi:10.2196/69678.
19. Martindale APL, Elhallak O, Pannell J, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. Nat Commun. 2024;15. doi:10.1038/s41467-024-45355-3.

Más allá del benchmark: cómo validar la IA con desenlaces clínicos reales

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Revisión narrativa estructurada dirigida a médicos de la Academia Nacional de Medicina de México

Síntesis estratégica dirigida a los miembros de la Academia Nacional de Medicina de México sobre la validación de la Inteligencia Artificial (IA) en la práctica médica. Bajo el liderazgo de los Dres. Rodolfo Palencia Díaz y Rodolfo de J Palencia Vizcarra (fundadores de TICC PalencIA), y en colaboración con el Dr. Raúl Carrillo Esper (presidente de la ANMM), se analiza la transición crítica del desempeño técnico algorítmico hacia la demostración de beneficios clínicos tangibles.

Contexto Estratégico: La Insuficiencia del Benchmark Técnico

La expansión de la IA médica ha centrado el debate inicial en métricas de rendimiento como el AUROC, la sensibilidad, la especificidad y la exactitud (F1). Sin embargo, estas métricas representan solo una capa superficial del problema. En la medicina real, el valor de una herramienta no reside en su capacidad de clasificación en un conjunto de datos estático, sino en su capacidad para:

- Cambiar decisiones clínicas de forma segura.
- Reducir errores médicos significativos.
- Evitar daños al paciente.
- Mantener la utilidad en diversos entornos institucionales y periodos temporales.

Para la medicina en México, esta distinción es fundamental. Un modelo validado exclusivamente en entornos altamente digitalizados corre el riesgo de degradar su desempeño al enfrentarse a la epidemiología operativa y la arquitectura asistencial de los hospitales locales.

La Escalera de Validación Clínica Real

La validación robusta de la IA debe seguir una secuencia escalonada que desplace el enfoque de la capacidad computacional hacia el beneficio clínico neto.

Niveles de Evaluación y Riesgos Asociados

Nivel de Evaluación	Pregunta Dominante	Medidas Principales	Riesgo si se Omite
1. Benchmark técnico	¿Clasifica bien en datos históricos?	AUC, sensibilidad, especificidad, F1, calibración	Confundir desempeño analítico con utilidad clínica.
2. Validación externa	¿Mantiene desempeño en otros hospitales y poblaciones?	Discriminación, calibración, análisis por sitio y subgrupo	Sobreajuste y falsa generalización.
3. Evaluación prospectiva silenciosa	¿Qué haría el sistema en tiempo real sin intervenir?	Concordancia, *drift*, errores de integración	Implementar sin conocer fallas operativas.
4. Evaluación clínica activa	¿Modifica decisiones y procesos de forma segura?	Tiempo a intervención, conducta clínica, eventos adversos	Daño por interacción humano-IA no estudiada.
5. Resultados del mundo real	¿Mejora lo que importa al paciente y al sistema?	Mortalidad, complicaciones, rehospitalización, PROM/PRO, costos	Adopción basada solo en variables sustitutas.

Síntesis de la Evidencia Actual

Las revisiones de ensayos controlados aleatorizados (ECA) muestran un patrón de resultados positivos concentrados en desenlaces de rendimiento (diagnóstico y procesos), pero con una marcada heterogeneidad en desenlaces clínicos duros.

Estudios Clave en la Transición Clínica

- Lam et al. (2022): Revisión de 39 ECA que muestran una señal favorable para la IA, pero con una generalización limitada debido al tamaño de las muestras y su naturaleza unicéntrica.
- Zhou et al. (2021): Revela que aproximadamente el 40% de las intervenciones evaluadas no ofrecieron un beneficio clínico claro frente a la atención estándar, subrayando que el benchmark no sustituye la prueba clínica.
- Han et al. (2024): En una revisión de 86 ECA, el 81% de los desenlaces primarios fueron positivos, pero centrados predominantemente en rendimiento diagnóstico y procesos de atención.

- Gommers et al. (2026): Representa un ejemplo sólido de validación avanzada en tamizaje mamográfico, demostrando menor tasa de cáncer de intervalo y mayor sensibilidad con IA.

La Brecha de la Centralidad del Paciente

Existe una deficiencia crítica en el uso de Medidas de Resultados Reportados por los Pacientes (PROM/PRO). Según Pearce et al. (2023), estas medidas se utilizan poco en los ensayos de IA, lo que refleja una evaluación todavía centrada en el algoritmo y no en el individuo.

Marcos de Transparencia y Control Metodológico

La validación clínica debe adherirse a marcos explícitos para garantizar la reproducibilidad y la seguridad:

- Diseño y Reporte: PRISMA 2020, AMSTAR-2, STARD-AI.
- Protocolos y Ensayos: CONSORT-AI, SPIRIT-AI, DECIDE-AI.
- Modelos de Predicción: TRIPOD+AI, PROBAST+AI.
- IA Generativa: GAMER.
- Confianza e Implementación: FUTURE-AI.

Algoritmo Propuesto para la Validación Clínica de IA

Para asegurar que la IA se convierta en una medicina útil y no solo en una promesa tecnológica, se propone el siguiente algoritmo de implementación:

1. Definir el problema: Identificar un desenlace clínico que realmente importe al paciente o al sistema de salud.
2. Benchmark técnico: Establecer un rendimiento aceptable con calibración explícita.
3. Validación externa: Realizar pruebas multicéntricas y por subgrupos.
4. Prueba silenciosa (Silent Trial): Evaluar el modelo en tiempo real sin influir en la atención para detectar fallas de integración.
5. Evaluación activa: Realizar estudios con diseño pragmático o bajo el marco DECIDE-AI.
6. Medición multidimensional: Incluir seguridad, carga laboral, conducta clínica, PROM/PRO y costos.
7. Escalamiento y Vigilancia: Monitoreo posimplementación para detectar drift (degradación del modelo) y auditoría de sesgos.

Conclusiones para la Academia Nacional de Medicina

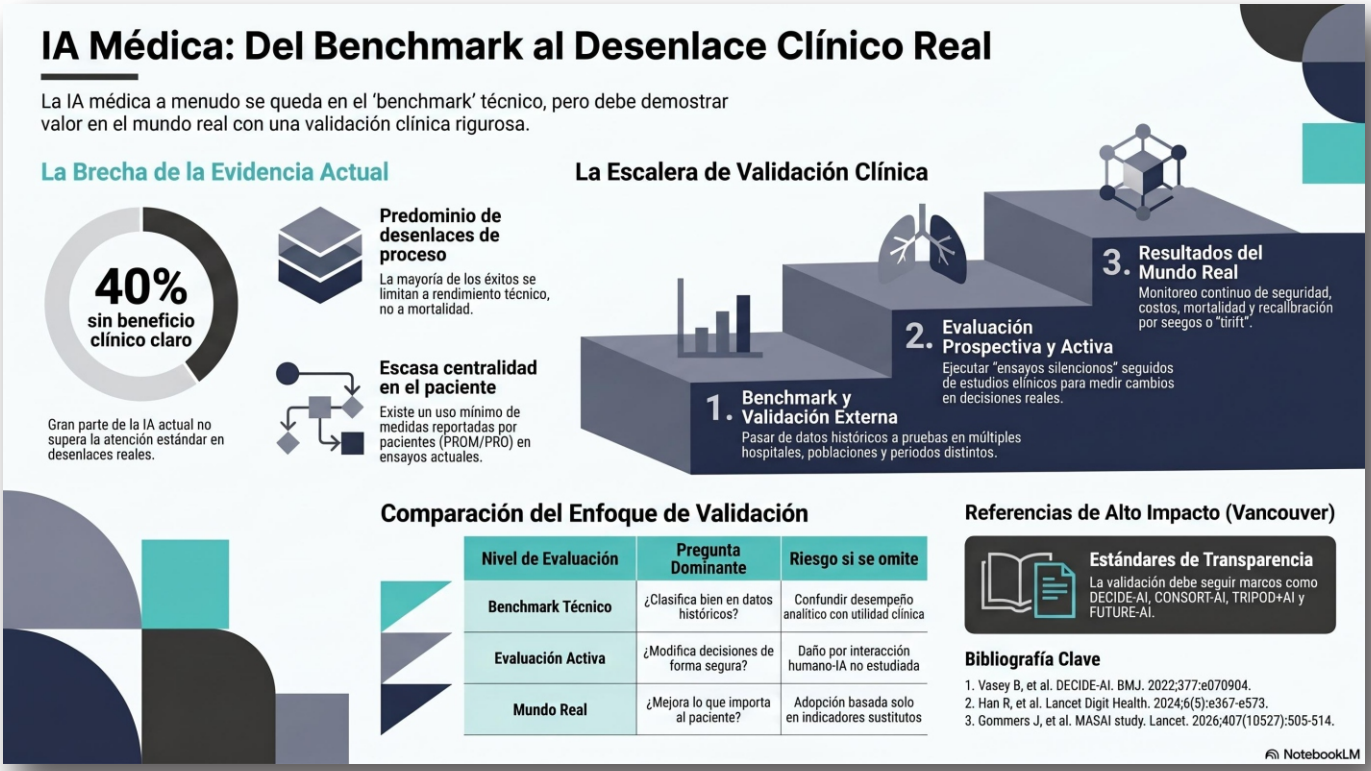
El benchmark es el inicio de la conversación, no su conclusión. Una herramienta de IA no debe escalarse si no demuestra un beneficio neto o, al menos, no inferioridad con ventajas operativas claras y seguridad mantenida.

La prioridad estratégica para los médicos de la Academia no es solo el desarrollo de nuevos modelos, sino la exigencia de estudios de mayor calidad: multicéntricos, transparentes, comparativos y centrados en desenlaces reales. La IA solo es medicina útil cuando resiste la prueba del paciente real y del flujo asistencial cotidiano.

Referencias Bibliográficas

1. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
2. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008. doi:10.1136/bmj.j4008.
3. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung DL, Sung JJY. Randomized controlled trials of artificial intelligence in clinical practice: systematic review. J Med Internet Res. 2022;24(8):e37188. doi:10.2196/37188.
4. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. npj Digit Med. 2021;4:154. doi:10.1038/s41746-021-00524-2.
5. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health. 2024;6(5):e367-e373. doi:10.1016/S2589-7500(24)00047-5.
6. Pearce FJ, Cruz Rivera S, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997-2022). Lancet Digit Health. 2023;5(3):e160-e167. doi:10.1016/S2589-7500(22)00249-7.
7. Khan SD, Ross M, Kocaballi AB, Magrabi F, Coiera E. Frameworks for procurement, integration, monitoring, and evaluation of artificial intelligence tools in clinical settings: a systematic review. PLOS Digit Health. 2024;3(5):e0000514. doi:10.1371/journal.pdig.0000514.

8. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. doi:10.1136/bmj-2022-070904.
9. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ. 2020;370:m3164. doi:10.1136/bmj.m3164.
10. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7.
11. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
12. Moons KGM, Dhiman P, Collins GS, Damen JAAG, Beam AL, Van Calster B, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
13. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-3289. doi:10.1038/s41591-025-03953-8.
14. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, et al. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ Evid Based Med. 2025;30(6):390- 400. doi:10.1136/bmjebm-2025-113825.
15. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. doi:10.1136/bmj-2024-081554.
16. Gommers J, Zackrisson S, Lång K, Andersson I, Timberg P, Strand F, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet. 2026;407(10527):505-514. doi:10.1016/S0140-6736(25)02464-X.



Más allá del benchmark: Cómo validar la IA con desenlaces clínicos reales

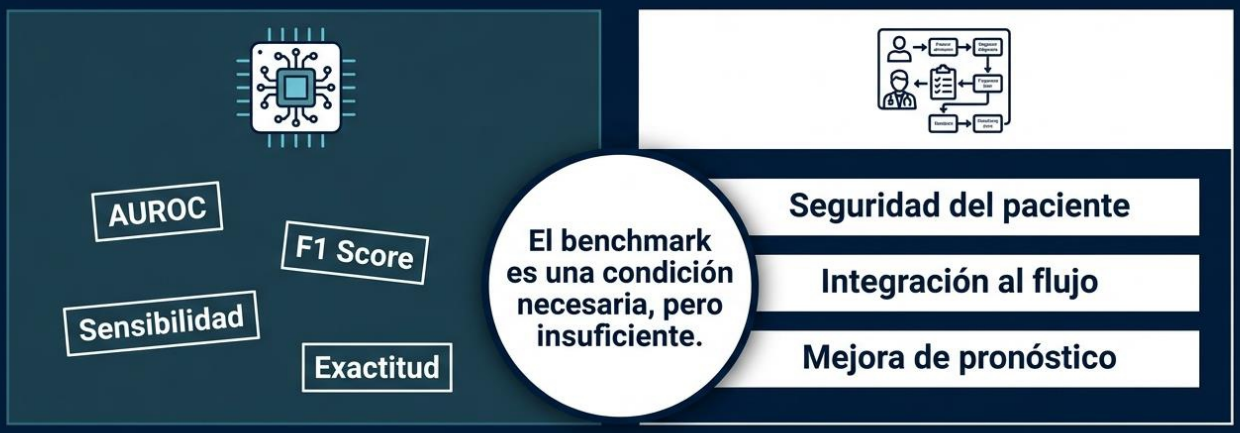
Hacia un nuevo estándar de rigor tecnológico centrado en el paciente.

Documento de revisión estructurada dirigido al Dr. Raúl Carrillo Esper y la Academia Nacional de Medicina de México (ANMM).

Dr. Rodolfo Palencia Díaz &
Dr. Rodolfo de J Palencia Vizcarra
Médicos Internistas, Fundadores de TICC PalencIA
Colegiados (CMIM) y Certificados (CMMI)

NotebookLM

Un sistema con buena discriminación retrospectiva fracasará sin validación en el flujo clínico real



La validación médica no puede detenerse en la capacidad computacional; debe demostrar beneficio clínico neto, seguridad sostenida y generalización operativa.

NotebookLM

La evidencia actual revela una concentración alarmante de resultados en desenlaces intermedios

Han et al., 2024	Zhou et al., 2021	Lam et al., 2022
81% desenlaces primarios positivos de 86 ECA. Alerta: Concentrados excesivamente en rendimiento diagnóstico o procesos.	**~40%** de intervenciones con IA **NO** ofrecen un beneficio clínico claro frente al estándar de atención.	**39 ECA** con señal favorable, pero severamente limitados por diseños pequeños y unicéntricos.

NotebookLM

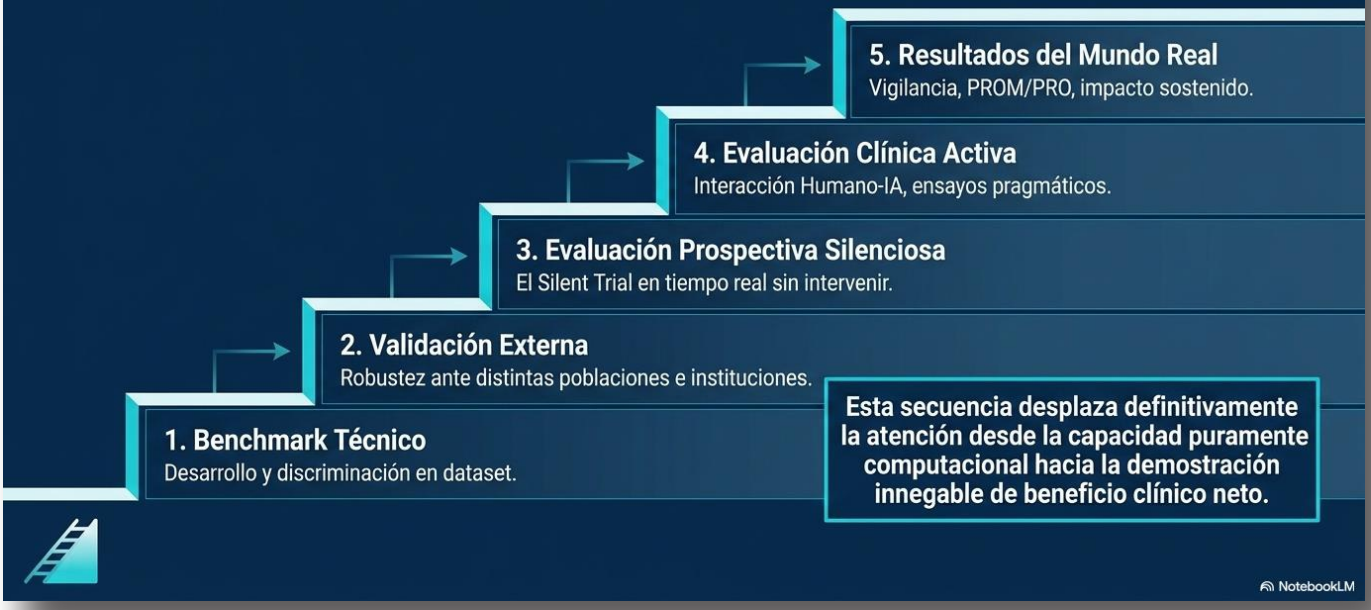
Matriz de Valor Translacional: El contraste entre la ilusión analítica y el impacto médico

	Nivel Benchmark	Nivel Clínico
Pregunta Dominante	¿Clasifica bien en datos históricos o similares del pasado?	¿Mejora lo que verdaderamente importa al paciente y al sistema hoy?
Métricas Principales	AUC, Sensibilidad, Especificidad, F1, Calibración.	Mortalidad, Complicaciones, Rehospitalización, PROM/PRO, Costos.
Riesgo Principal del Enfoque	Confundir la capacidad analítica aislada con utilidad médica real.	Adopción tecnológica basada exclusivamente en marcadores subrogados.

Síntesis conceptual basada en las limitaciones detectadas en los marcos DECIDE-AI y FUTURE-AI.

NotebookLM

La Escalera de Validación Clínica exige un tránsito metodológico escalonado



NotebookLM

Un modelo encajado en su entorno original degrada su precisión al cambiar de contexto

Overfitting vs. Generalization

Hospital A - Entorno de Origen

95% AUC. Trazabilidad perfecta, epidemiología homogénea.

Hospital B - El Mundo Real

Degradación severa del desempeño. Diferente arquitectura asistencial, variables faltantes, nueva epidemiología.

La **Validación Externa** es innegociable. Asumir generalización sin probarla en múltiples hospitales, equipos y periodos temporales conduce a falsas promesas de seguridad.

Evaluación Prospectiva Silenciosa: Detectando fallas operativas con riesgo cero para el paciente

Parallel Tracks

Top Track - Vía A - Flujo Estándar

Atención clínica habitual continúa sin interrupciones.

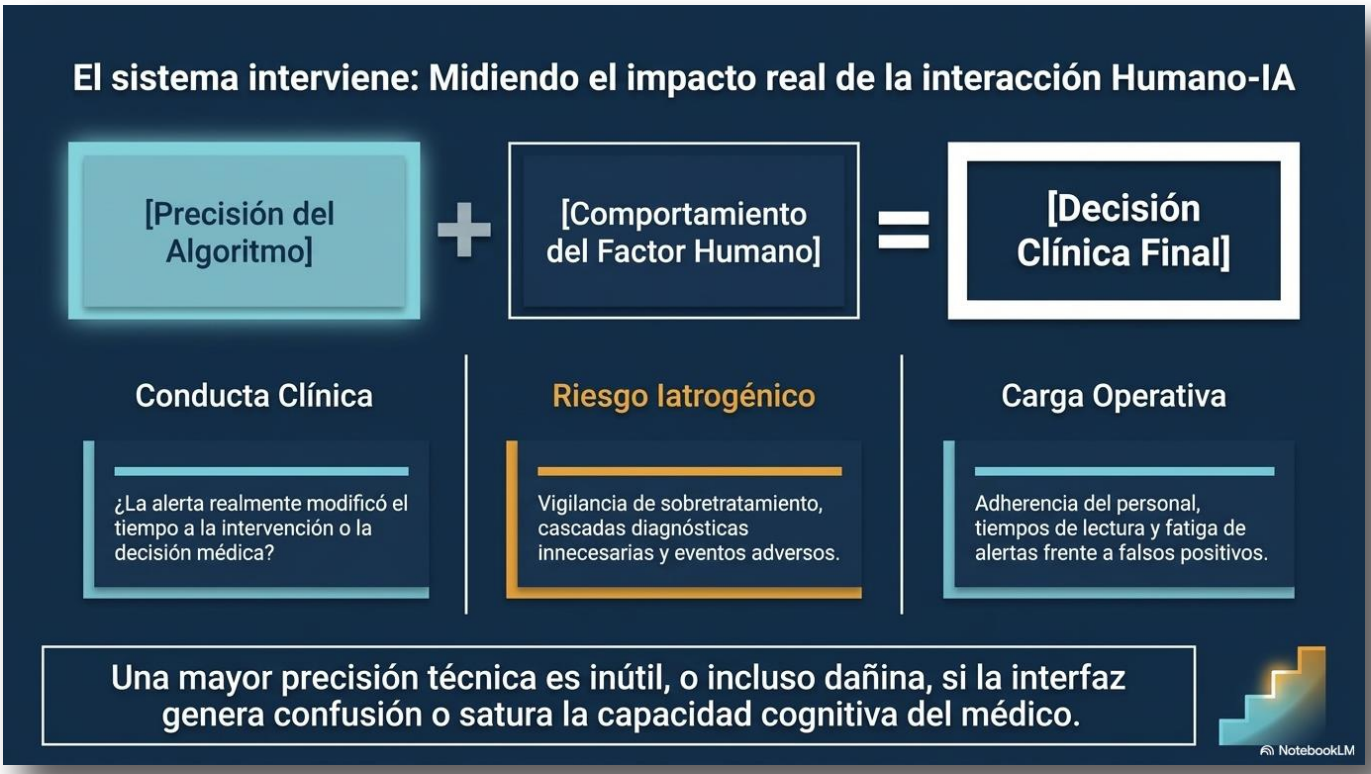
Comparación de Concordancia

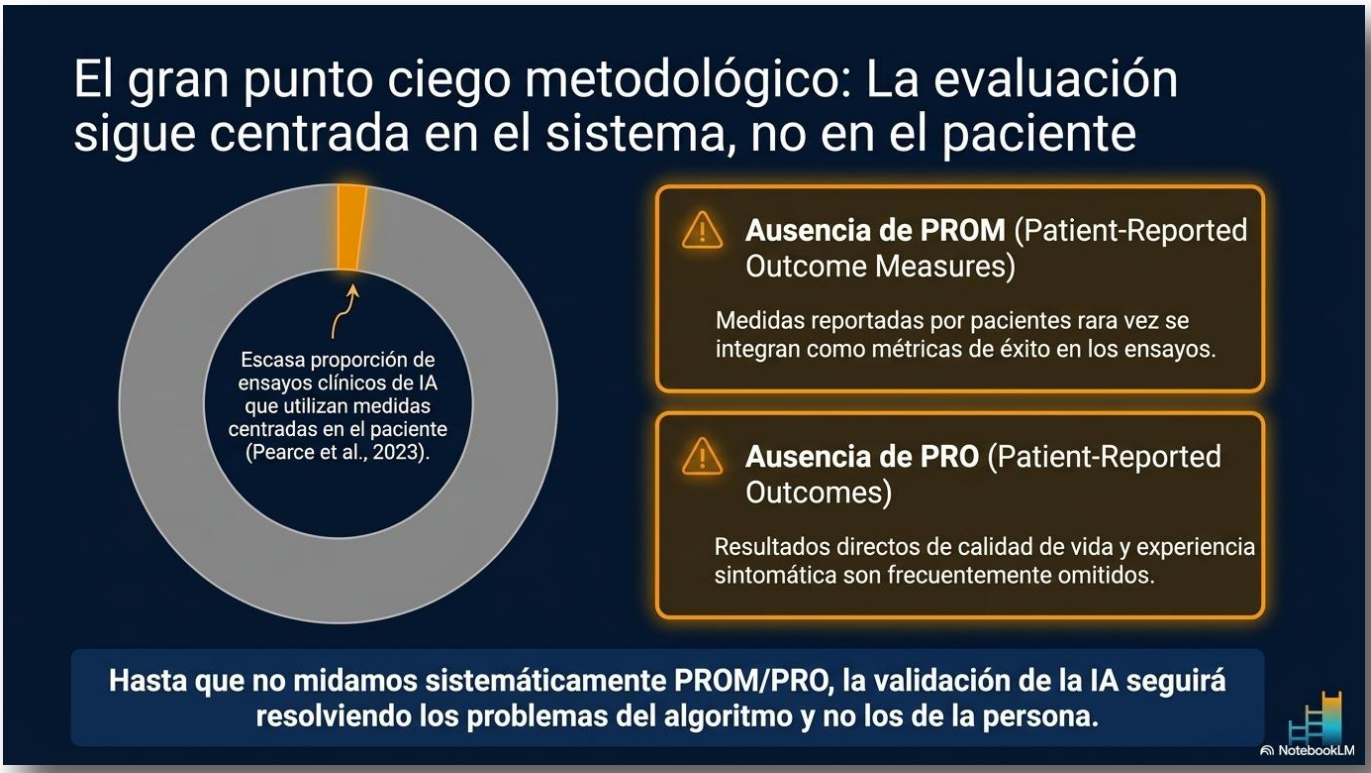
Bottom Track - Vía B - La IA en la Sombra

Evaluación no intervencionista en tiempo real.

¿Qué medimos aquí?

- Drift algorítmico en datos reales.
- Errores de integración al sistema hospitalario.
- Carga potencial de fatiga por alertas excesivas.


El sistema interviene: Midiendo el impacto real de la interacción Humano-IA
[Precisión del Algoritmo]
+
[Comportamiento del Factor Humano]
=
[Decisión Clínica Final]
Conducta Clínica
¿La alerta realmente modificó el tiempo a la intervención o la decisión médica?
Riesgo Iatrogénico
Vigilancia de sobretratamiento, cascadas diagnósticas innecesarias y eventos adversos.
Carga Operativa
Adherencia del personal, tiempos de lectura y fatiga de alertas frente a falsos positivos.
Una mayor precisión técnica es inútil, o incluso dañina, si la interfaz genera confusión o satura la capacidad cognitiva del médico.
NotebookLM


El gran punto ciego metodológico: La evaluación sigue centrada en el sistema, no en el paciente
Escasa proporción de ensayos clínicos de IA que utilizan medidas centradas en el paciente (Pearce et al., 2023).
Ausencia de PROM (Patient-Reported Outcome Measures)
Medidas reportadas por pacientes rara vez se integran como métricas de éxito en los ensayos.
Ausencia de PRO (Patient-Reported Outcomes)
Resultados directos de calidad de vida y experiencia sintomática son frecuentemente omitidos.
Hasta que no midamos sistemáticamente PROM/PRO, la validación de la IA seguirá resolviendo los problemas del algoritmo y no los de la persona.
NotebookLM

Un Estándar de Oro Metodológico: El Ensayo Clínico MASAI

Gommers et al., 2026: ECA poblacional de no inferioridad en tamizaje mamográfico asistido por IA.

Eficacia Oncológica

Menor tasa comprobada de cáncer de intervalo.

Mantenimiento de Calidad

Sensibilidad y especificidad sostenidas frente al estándar humano.

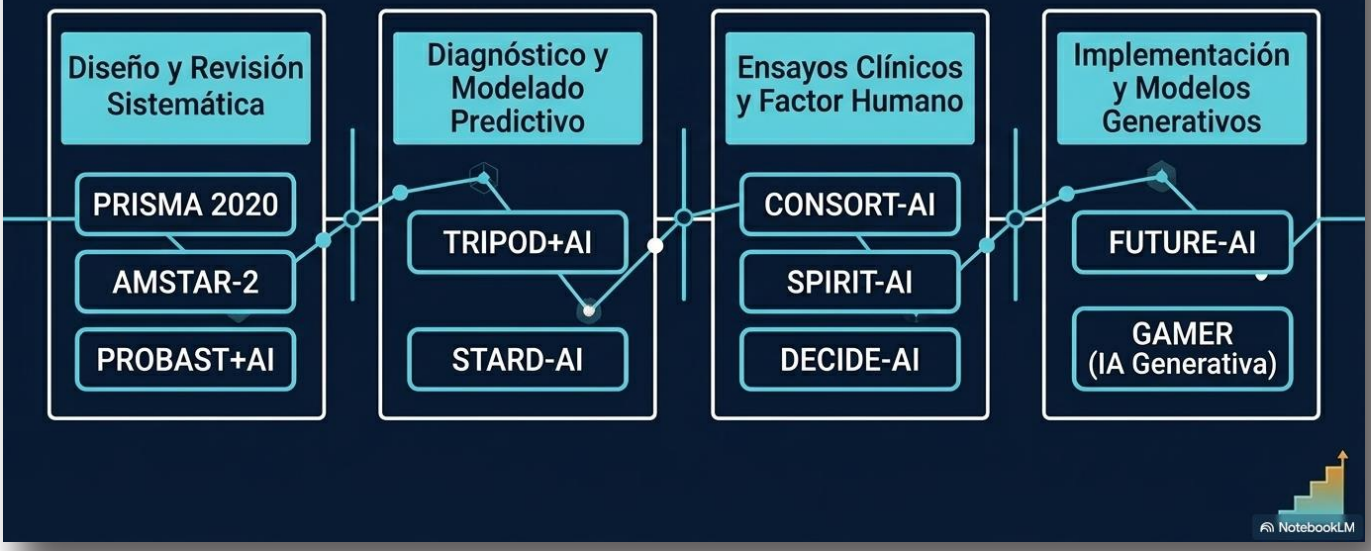
Beneficio Operativo

Reducción estadísticamente significativa en la carga de lectura humana.

Este estudio representa el estándar exigible: no se conformó con probar que la IA "veía bien" la imagen en un dataset retrospectivo; documentó un beneficio neto en la arquitectura asistencial protegiendo la seguridad poblacional.

NotebookLM

Navegando el ecosistema normativo: Marcos de control metodológico por fase



NotebookLM


El Reto para México: Transferibilidad e incertidumbres en nuestro ecosistema de salud
IA Extranjera
Brecha de Madurez Digital
La literatura de calidad proviene de sistemas con trazabilidad y digitalización radicalmente distintas a la de los hospitales públicos en América Latina.
Vacíos de Especialidad
La evidencia robusta se concentra en imagen y detección, pero es débil en medicina interna general, urgencias y el manejo de multimorbilidad compleja.
Sesgo Epidemiológico
Asumir transferibilidad automática sin validación multicéntrica local constituye constituye un riesgo metodológico grave.
Imperativo Estratégico: México necesita impulsar validación contextualizada e independiente antes del despliegue masivo.
NotebookLM


Blueprint Metodológico: Algoritmo operativo de validación clínica en 7 pasos
1. Definir el Problema
Identificar un desenlace que importe genuinamente al paciente o al sistema.
2. Benchmark con Calibración
Demostrar desempeño técnico aceptable con análisis de error explícito.
3. Validación Externa
Prueba multicéntrica en subgrupos relevantes.
4. Silent Trial
Evaluación prospectiva silenciosa (detección de drift y fallas de integración).
5. Evaluación Activa
Ensayos pragmáticos (DECIDE-AI) si la señal previa es favorable.
6. Medición Integral
Evaluar tiempos, conducta médica, carga laboral, seguridad y PROM/PRO.
7. Escalamiento y Auditoría
Monitoreo posimplementación, recalibración continua y auditoría de sesgos.
NotebookLM

El benchmark es el inicio de la conversación científica, no su punto final

Una herramienta no debe escalarse al nivel asistencial si no demuestra beneficio neto, o al menos, no inferioridad con ventajas operativas claras y seguridad sostenida.

> Para la Academia Nacional de Medicina de México, la prioridad estratégica **no es desarrollar más algoritmos, sino exigir mejores estudios: multicéntricos, pragmáticos, transparentes y centrados en la vida del paciente.**

NotebookLM

Referencias Bibliográficas

Page MJ, et al. The PRISMA 2020 statement... BMJ. 2021;372:n71.

Shea BJ, et al. AMSTAR 2... BMJ. 2017;358:j4008.

Lam TYT, et al. Randomized controlled trials of artificial intelligence... J Med Internet Res. 2022.

Zhou Q, et al. Clinical impact and quality of randomized controlled trials... npj Digit Med. 2021.

Han R, et al. Randomised controlled trials evaluating artificial intelligence... Lancet Digit Health. 2024.

Pearce FJ, et al. The role of patient-reported outcome measures... Lancet Digit Health. 2023.

Khan SD, et al. Frameworks for procurement, integration, monitoring... PLOS Digit Health. 2024.

Vasey B, et al. Reporting guideline for early-stage clinical evaluation: DECIDE-AI. BMJ. 2022.

Liu X, et al. The CONSORT-AI extension. BMJ. 2020.

Cruz Rivera S, et al. The SPIRIT-AI extension. Nat Med. 2020.

Collins GS, et al. TRIPOD+AI statement... BMJ. 2024.

Moons KGM, et al. PROBAST+AI... BMJ. 2025.

Sounderajah V, et al. The STARD-AI reporting guideline... Nat Med. 2025.

Luo X, et al. The GAMER Statement. BMJ Evid Based Med. 2025.

Lekadir K, et al. FUTURE-AI... BMJ. 2025.

Gommers J, et al. Interval cancer, sensitivity, and specificity... MASAI study... Lancet. 2026.

NotebookLM

Gobernanza de la IA en salud: qué deberían exigir los hospitales mexicanos antes de comprarla o usarla

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Revisión narrativa estructurada dirigida a médicos de la Academia Nacional de Medicina de México

Este documento sintetiza la propuesta metodológica y los controles críticos presentados por los Dres. Rodolfo Palencia Díaz y Rodolfo de J. Palencia Vizcarra (TICC PalencIA), en colaboración con el Dr. Raúl Carrillo Esper (Presidente de la Academia Nacional de Medicina de México). El objetivo es guiar a las instituciones sanitarias mexicanas en la adopción responsable de la Inteligencia Artificial (IA), trascendiendo la lógica comercial para priorizar la seguridad clínica y la gobernanza institucional.

Mensaje Central

Un hospital no debe adquirir IA como un software genérico. La institución debe exigir un paquete verificable de evidencia clínica, controles de riesgo, cumplimiento regulatorio, trazabilidad contractual y vigilancia continua. La utilidad de una herramienta no reside en su sofisticación técnica, sino en su capacidad de ser gobernada dentro del entorno clínico y jurídico mexicano.

1. El Cambio de Paradigma: De la Métrica Técnica a la Gobernanza Clínica

La evaluación tradicional de la IA basada en sensibilidad, especificidad o el área bajo la curva (AUC) es necesaria pero insuficiente. La gobernanza de la IA debe tratarse como una función clínica y organizacional distribuida en seis capas fundamentales:

- Clínica
- Metodológica
- Jurídica
- Técnica
- Operativa
- Contractual

Ciclo de Gobernanza Hospitalaria Continua

El valor de una herramienta debe reexaminarse a lo largo de todo su ciclo de vida:

1. Definir el caso de uso.
2. Clasificar el riesgo y la regulación.
3. Auditar la evidencia.
4. Validar localmente.
5. Contratar con controles.
6. Desplegar de forma escalonada.
7. Monitorear, corregir o retirar.

2. Las Seis Exigencias Críticas para Instituciones de Salud

De acuerdo con la evidencia y los marcos internacionales, los hospitales deben imponer los siguientes requisitos mínimos:

Tabla 1. Requerimientos no negociables y señales de alerta

Dominio	Qué debe exigir el hospital	Señales de alerta (Red Flags)
Propósito Clínico	Caso de uso preciso, población objetivo y decisión que apoya.	Promesa genérica de "mejora de productividad".
Clasificación Regulatoria	Definir si es apoyo administrativo o software con propósito médico.	El proveedor evita definir su estatus regulatorio.
Evidencia Clínica	Validación externa, calibración y errores críticos desglosados.	Solo muestran benchmarks internos o AUC simple.
Datos y Privacidad	Origen de datos, base jurídica, cifrado y derecho a auditoría.	No especifican dónde ni cómo se procesan los datos.
Sesgo y Equidad	Desempeño analizado por subgrupos poblacionales relevantes.	Afirman "ausencia de sesgo" sin análisis estratificado.
Control de Cambios	Plan de cambios, control de versiones y capacidad de *rollback*	Actualizaciones automáticas y o pacas.

3. Marco Regulatorio y Normativo en México

La adopción de IA en México debe alinearse con el ecosistema legal vigente y las reformas recientes:

- Ley General de Salud (Reforma 2026): Incorpora el capítulo de Salud Digital, obligando a contar con infraestructura, protocolos de seguridad y mecanismos de evaluación institucional.
- NOM-024-SSA3-2012: Crucial para la interoperabilidad con el Expediente Clínico Electrónico.
- NOM-241-SSA1-2025: Aplicable si la IA opera como dispositivo médico o software con finalidad médica (SaMD).
- Protección de Datos: Cumplimiento estricto de la *Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados y la Ley Federal de Protección de Datos Personales en Posesión de los Particulares,* especialmente en lo referente al uso secundario de datos para entrenamiento algorítmico.

4. Estándares Metodológicos para la Auditoría de Evidencia

El comité hospitalario no debe aceptar revisiones superficiales. Se debe exigir el cumplimiento de estándares internacionales según el tipo de herramienta:

Tabla 2. Estándares metodológicos recomendados

Estándar	Utilidad para el Comité Hospitalario
DECIDE-AI	Valorar si el piloto clínico real fue reportado con rigor.
CONSORT-AI / SPIRIT-AI	Asegurar que los ensayos clínicos y protocolos tengan un diseño adecuado para IA.
STARD-AI	Interpretar sesgos y precisión diagnóstica en el flujo de pacientes.
TRIPOD+AI	Evaluar la validez y generalización de modelos predictivos.
CHEERS-AI	Validar si el ahorro económico prometido está justificado metodológicamente.
GAMER / RAISE	Declarar el uso de IA generativa y evitar la automatización opaca en dictámenes.

5. Algoritmo Práctico para el Despliegue de IA

Para asegurar una transición segura, se propone la siguiente secuencia de decisión institucional:

1. Definir el problema: Identificar la necesidad clínica u operativa real.
2. Delimitar el caso de uso: Establecer qué hace, qué NO hace y quién conserva la decisión final (supervisión humana).
3. Clasificación técnica: Categorizar como herramienta administrativa, de apoyo o médica.
4. Criba jurídico-regulatoria: Verificar cumplimiento con LGS 2026, NOMs y COFEPRIS.
5. Auditoría de evidencia: Revisar validación externa y errores clínicamente relevantes.
6. Piloto local controlado: Evaluar seguridad, flujo de trabajo y carga cognitiva del médico.
7. Contrato y monitoreo: Establecer cláusulas de auditoría, rollback y monitoreo de drift (deterioro del modelo).

6. Conclusiones y Recomendaciones para la ANMM

La IA sanitaria debe ser tratada como una intervención institucional de alto control. La pregunta estratégica para los miembros de la Academia Nacional de Medicina de México no es qué tan inteligente parece la herramienta, sino qué tan gobernable es dentro de su hospital.

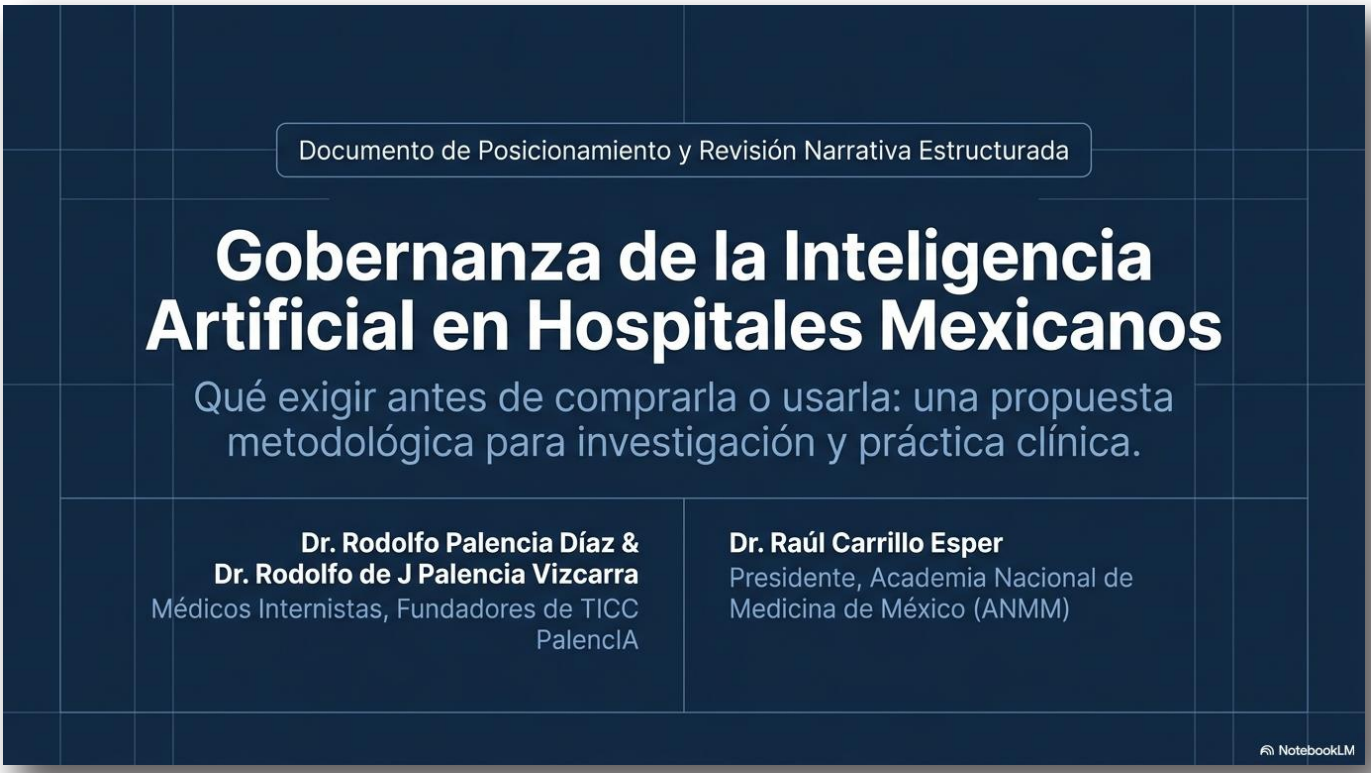
Puntos clave finales:

- La validación externa no sustituye la validación local.
- El contrato debe garantizar el derecho a auditoría y la causal de retiro ante incidentes.
- La gobernanza robusta no frena la innovación; la hace defendible ante la ley y segura para el paciente.

Referencias Bibliográficas

1. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021.
2. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2025.
3. The DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9.
4. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x.
5. México. Ley General de Salud. Reforma publicada en el Diario Oficial de la Federación el 15 de enero de 2026. Capítulo VI Bis, Salud Digital.

6. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. Silver Spring, MD: FDA; 2024.
7. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. doi:10.1136/bmj-2024-081554.
8. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical predictionmodels that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
9. México. NORMA Oficial Mexicana NOM-241-SSA1-2025, Buenas prácticas de fabricación de dispositivos médicos. Diario Oficial de la Federación, 4 de abril de 2025.
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.


Documento de Posicionamiento y Revisión Narrativa Estructurada
Gobernanza de la Inteligencia Artificial en Hospitales Mexicanos
Qué exigir antes de comprarla o usarla: una propuesta metodológica para investigación y práctica clínica.
Dr. Rodolfo Palencia Díaz &
Dr. Rodolfo de J Palencia Vizcarra
Médicos Internistas, Fundadores de TICC PalencIA
Dr. Raúl Carrillo Esper
Presidente, Academia Nacional de Medicina de México (ANMM)
NotebookLM


Un hospital no debería comprar IA como si adquiriera un software genérico. Debe tratarse como una intervención institucional de alto control.
Promesa Tecnológica
La literatura crece rápidamente en validación retrospectiva y métricas de desempeño, pero falla sistemáticamente en evaluar el contexto real y los desenlaces clínicos en pacientes.
Realidad Clínica
La prioridad estratégica no es qué tan "inteligente" o precisa (AUC) parece la herramienta, sino qué tan gobernable es dentro del entorno clínico, jurídico y operativo del hospital.
NotebookLM

Cambio de Paradigma

Adquisición TI Tradicional	Gobernanza de IA en Salud
Objeto de compra: Licencia de software informático.	**Objeto de compra:** Intervención clínica u operativa delegada.
Métrica de éxito: Mejora de productividad y despliegue rápido.	**Métrica de éxito:** Seguridad, equidad y desenlaces clínicos medibles.
Validación requerida: Benchmark interno y métricas del proveedor.	**Validación requerida:** Validación externa, calibración local y piloto en flujo de trabajo.
Fin del ciclo: Instalación funcional y pago de licencia.	**Fin del ciclo:** Monitoreo continuo, control de cambios y causal de retiro predefinida.

Comprar IA sin una arquitectura de gobernanza equivale a desplazar el riesgo desde el proveedor hacia la institución y hacia el médico responsable.

NotebookLM

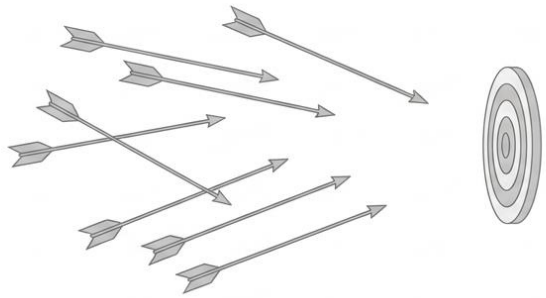
El Ecosistema de Gobernanza

El ciclo demuestra que el valor y la seguridad de una herramienta no terminan en el contrato comercial; la inteligencia artificial en salud exige reexaminarse perpetuamente a lo largo de toda su vida útil.

NotebookLM

Exigencia 1: Definir la Indicación Clínica u Operativa

Bandera Roja



"Sirve para múltiples áreas" o "Mejora la productividad general".

Riesgo: Sin una tarea definida, cualquier comparación metodológica se vuelve inválida.

Gobernanza

El comité hospitalario debe exigir una Ficha de Indicación estricta antes de evaluar cualquier sistema:

- ☐ Problema clínico u operativo exacto.
- ☐ Población objetivo estrictamente delimitada.
- ☐ Insumos requeridos.
- ☐ Tipo de usuario final y límites de automatización.
- ☐ Qué decisión apoya (y qué decisión NO apoya).

NotebookLM

Exigencia 2: Clasificación y Ruta Regulatoria en México

¿Qué hace la herramienta IA?

- **Ruta A** (Gestión Administrativa) → **Cumplimiento de Privacidad de Datos** (LGPDPPSO / LFPDPPP)
- **Ruta B** (Conectada a Expediente / Telesalud) → **Interoperabilidad y Seguridad de la Información** (NOM-024-SSA3-2012)
- **Ruta C** (Propósito Médico / Apoyo a Decisión) → **Software como Dispositivo Médico (SaMD)** → Intervención de COFEPRIS y Buenas Prácticas de Fabricación (NOM-241-SSA1-2025)

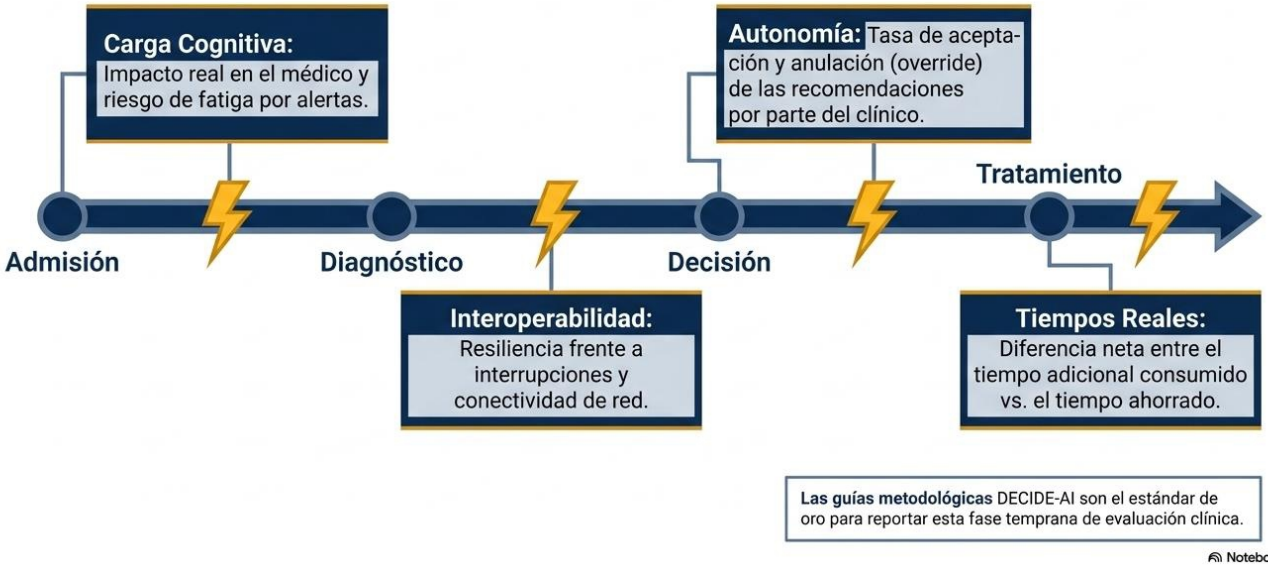
La **reforma a la Ley General de Salud (Capítulo Salud Digital, 2026)** obliga a los hospitales a prever infraestructura, capacitación, protocolos de seguridad y mecanismos formales de evaluación para estas tecnologías.

NotebookLM

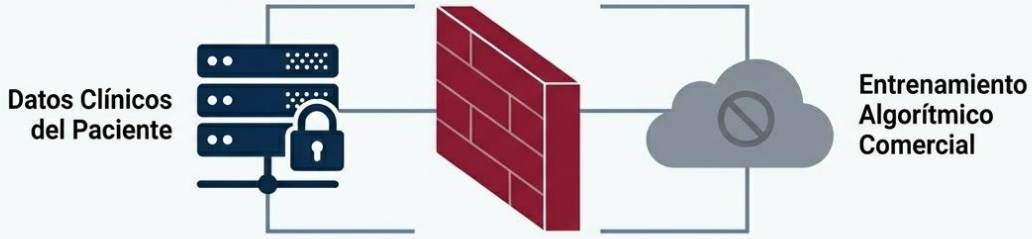
Exigencia 3: Evidencia Clínicamente Útil (Más allá del AUC)

Exigencia 4: Validación en el Flujo de Trabajo Real

Existe una brecha persistente entre el rendimiento in vitro de un algoritmo y su integración clínica. Exige un piloto local controlado antes del despliegue masivo.



Exigencia 5: Gobernanza de Privacidad y Consentimiento



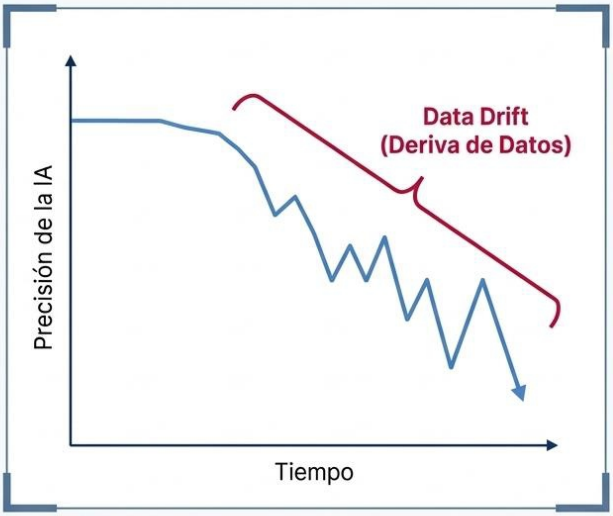
Sector Público: Ley General (LGPDPPSO) → Exige evaluación de impacto para tratamiento intensivo o sensible.

Sector Privado: Ley Federal (LFPDPPP) → Principios de licitud, finalidad, proporcionalidad y responsabilidad.

Regla de Oro Contractual:
La base jurídica del tratamiento asistencial NUNCA debe confundirse con la base jurídica del entrenamiento algorítmico. Los hospitales deben prohibir contratos ambiguos que permitan al proveedor usar la información sensible de sus pacientes para "mejora del modelo", desarrollo posterior o reentrenamiento comercial sin anonimización auditable y consentimiento explícito.

Exigencia 6: Vigilancia y Control de Cambios del Modelo

La **evaluación precompra** no es suficiente. Los **algoritmos se deterioran** por **cambios epidemiológicos**, **drift** de datos y modificaciones del flujo clínico.



Garantías Contractuales Innegociables	
1.	**Derecho a Auditoría**: Acceso irrestricto a los logs del sistema y reporte de incidentes.
2.	**Notificación de Cambios**: Planes de control predeterminados (alineados a estándares FDA) para cada actualización del algoritmo.
3.	**Capacidad de Rollback**: Infraestructura para regresar inmediatamente a la versión anterior si una actualización falla.
4.	**Causal de Retiro**: Protocolo predefinido para "apagar" el algoritmo si pierde equidad, seguridad o utilidad.

Matriz de Diagnóstico: Señales de Alerta Comerciales

Síntoma Comercial (Lo que dice el vendedor)	Diagnóstico (El riesgo institucional)	Tratamiento (Lo que debe exigir el hospital)
La herramienta optimiza procesos y sirve para múltiples áreas.	Falta de delimitación. Riesgo de uso 'off-label'.	Exigir Ficha de Indicación precisa y exclusiva.
Tenemos un AUC del 98% en nuestro hospital de origen.	Ausencia de validación externa. Alto riesgo de sobreajuste poblacional.	Exigir análisis de errores críticos, calibración y piloto local.
Nuestros algoritmos aprenden y se actualizan solos para ser mejores.	Modificaciones opacas sin validación clínica de los cambios.	Exigir versiones controladas, umbrales de revalidación y protocolo rollback.
El servicio es en la nube; optimizamos con el uso del hospital.	Explotación secundaria de datos sensibles sin base legal.	Bloqueo contractual de transferencia y exigencia de auditoría.

Estándares Metodológicos para la Evidencia Clínica

DECIDE-AI
Para evaluación clínica temprana y pilotos en entornos de flujo de trabajo real.

CONSORT-AI / SPIRIT-AI
Para juzgar protocolos y resultados de ensayos clínicos aleatorizados.

STARD-AI
Para interpretar estudios de precisión diagnóstica y flujo de pacientes.

TRIPOD+AI
Para evaluar validez y generalización de modelos predictivos y machine learning.

CHEERS-AI
Para auditar evaluaciones económicas y corroborar ahorros justificados.

GAMER / RAISE
Para uso de IA Generativa en investigación y síntesis (obligando a declarar qué decisión siguió siendo humana).

Algoritmo de Decisión para Hospitales Mexicanos

Base Crítica: Suspender o retirar si aparecen drift, sesgo inaceptable, brechas de ciberseguridad o pérdida de trazabilidad.

NotebookLM

Checklist Ejecutivo para Comités Hospitalarios

☐	¿La herramienta tiene un caso de uso clínico delimitado y excluye usos no autorizados?
☐	¿Existe definición explícita del usuario final y del nivel de supervisión humana requerido?
☐	¿Se aclaró si la herramienta se clasifica como software con propósito médico (NOM-241)?
☐	¿La evidencia incluye validación externa, calibración y análisis de errores críticos?
☐	¿Se exige una validación local / piloto de flujo de trabajo antes del despliegue amplio?
☐	¿Se auditaron la base jurídica del tratamiento, las transferencias y la prohibición de uso secundario?
☐	¿El contrato otorga a la institución derecho de auditoría, control de cambios y rollback?
☐	¿Existen indicadores definidos para activar una causal de suspensión o retiro automático?

NotebookLM

"Para la Academia Nacional de Medicina de México, la gobernanza robusta no frena la innovación; la hace defendible. La pregunta estratégica de un hospital no es qué tan sofisticada parece la IA, sino qué tan gobernable es dentro de su realidad operativa."

Referencias Bibliográficas (Selección)

1. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021.
2. the DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933.
3. Khan SD, et al. Frameworks for procurement, integration, monitoring, and evaluation of artificial intelligence tools in clinical settings: a systematic review. PLOS Digit Health. 2024;3(5):e0000514.
4. México. Ley General de Salud. Reforma publicada DOF 15 enero 2026. Cap VI Bis, Salud Digital.
5. U.S. Food and Drug Administration. Transparency for machine learning-enabled medical devices: guiding principles. Silver Spring: FDA; 2024.

Competencias Mínimas en Inteligencia Artificial para el Médico Mexicano del Siglo XXI

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Revisión narrativa estructurada dirigida a médicos de la Academia Nacional de Medicina de México

Este documento sintetiza la propuesta técnica y académica para establecer un núcleo esencial de competencias en Inteligencia Artificial (IA) dirigido a los profesionales de la salud en México. Bajo la perspectiva de la medicina interna y la transformación digital, se define que la alfabetización en IA no es una destreza opcional de productividad, sino una competencia profesional de seguridad indispensable para la práctica clínica contemporánea.

Resumen

La integración de la IA en la medicina mexicana ha trasladado el debate educativo desde la pertinencia de su enseñanza hacia la definición de estándares mínimos que garanticen la seguridad del paciente, la privacidad de los datos y la integridad académica. La competencia mínima no reside en la habilidad técnica de programar o interactuar con un chatbot, sino en la capacidad de auditar salidas, comprender sesgos y sostener la decisión clínica final mediante el criterio humano y evidencia verificable.

Dominios Troncales de Competencia

Se proponen seis dominios esenciales que constituyen el "piso común" para médicos generales, especialistas, docentes y residentes.

Dominio	Conocimiento Indispensable	Habilidad Observable	Error Crítico a Evitar
1. Fundamentos de IA	Diferenciar IA predictiva, generativa y apoyo a la decisión. Comprender alucinación y deriva.	Explicar qué hace y qué no hace una herramienta en lenguaje clínico claro.	Confundir una respuesta convincente con una respuesta verdadera.
2. Evaluación Crítica	Conocer validez interna/externa, calibración, sesgo y vigilancia posdespliegue.	Preguntar por validación local, tasa de error y contexto de uso.	Adoptar herramientas por novedad o estrategias de marketing.
3. Datos y Privacidad	Reconocer datos identificables, minimización de datos y normatividad del expediente clínico.	Desidentificar y limitar la carga de datos a entornos autorizados.	Subir información sensible a sistemas no aprobados.
4. Ética y Comunicación	Comprender automatización acrítica, inequidad algorítmica y transparencia.	Informar al paciente o equipo cuándo la IA fungió como apoyo auxiliar.	Transferir responsabilidad moral o clínica a la herramienta.
5. Integración Clínica	Identificar tareas donde la IA ayuda sin desplazar el juicio humano.	Usar IA en tareas de bajo/mediano riesgo con verificación sistemática.	Copiar y pegar salidas no verificadas al expediente.
6. Investigación	Conocer guías internacionales (PRISMA, CONSORT-AI, GAMER, RAISE).	Declarar uso de IA, conservar trazabilidad y verificar referencias.	Presentar texto o referencias generadas como evidencia real.

Evidencia y Marcos de Referencia

La revisión estructurada de la literatura actual (2021-2026) señala que el médico del futuro es un profesional híbrido. Los marcos internacionales de consenso (como FUTURE-AI y DECODE) subrayan que la IA en salud debe ser equitativa, trazable, robusta y explicable.

Fuentes Clave y Mensajes Prácticos

- Schubert et al. (2025): Establece tres niveles de experticia; el nivel mínimo debe ser generalista y no técnico-profundo.
- Gazquez-Garcia et al. (2025): Identifica que los "fundamentos" se centran en juzgar precisión y límites de uso, no en programación.
- Wilhelm et al. (2025): Advierte que los beneficios y daños de los sistemas algorítmicos aún se miden de forma inconsistente, obligando a una vigilancia estrecha.
- Lekadir et al. (2025): Propone los principios de equidad y robustez como atributos mínimos que todo médico debe reconocer en una IA segura.

Traducción al Contexto Mexicano

En México, la adopción de IA está supeditada a un marco normativo estricto que incluye:

1. NOM-004-SSA3-2012: Relativa al expediente clínico.
2. NOM-024-SSA3-2012: Sobre sistemas de registro electrónico para la salud.
3. Lineamientos del INAI: Protección de datos personales en el sector público.
4. Guía Metodológica de la Secretaría de Salud (2025): Insiste en que la IA es una herramienta auxiliar y sus salidas deben ser validadas antes de cualquier empleo clínico.

Conductas Esperadas en el Entorno Hospitalario

Escenario	Conducta Esperada	Riesgo de Incumplimiento
Urgencias / Hospitalización	Verificar congruencia con guías y contexto clínico; uso solo como apoyo.	Automatización acrítica en pacientes de alto riesgo.
Documentación Clínica	No incorporar texto generado sin revisión integral del médico tratante.	Errores factuales e incongruencias en el expediente.
Docencia	Explicar límites y sesgos a residentes antes de usar la herramienta.	Aprendizaje superficial y dependencia tecnológica.
Investigación	Corroborar DOIs y afirmaciones; declarar uso de herramientas generativas.	Citas falsas y daño a la integridad académica.

Algoritmo Práctico para el Uso Clínico Responsable

Ante una necesidad clínica o académica, el facultativo debe seguir esta ruta lógica:

1. Definir la tarea: ¿Es búsqueda, resumen, borrador o apoyo documental?
2. Clasificar el riesgo: ¿Bajo, medio o alto?
3. Verificar entorno: ¿La herramienta está aprobada por la institución? ¿Es segura para los datos?
4. Minimizar datos: Desidentificar y no cargar información innecesaria.
5. Auditar la salida: ¿Es factual? ¿Cita fuentes verificables? ¿Es congruente con el paciente?
6. Decidir conducta: Ante discordancia o alto riesgo, no delegar y escalar a revisión humana.
7. Documentar: Registrar la decisión final humana; monitorizar errores.

Conclusiones y Agenda para la ANMM

La competencia en IA para el médico mexicano no consiste en perseguir la novedad tecnológica, sino en poseer una alfabetización crítica que preserve el juicio clínico. Para la Academia Nacional de Medicina de México, se sugieren tres niveles de acción:

- Doctrinal: Emitir una postura nacional que subordine la IA al juicio clínico y al interés del paciente.
- Curricular: Definir un módulo común de alfabetización en IA con evaluación de competencias observables.
- Institucional: Exigir políticas explícitas sobre el uso de herramientas, supervisión y reporte de errores.

Las 5 Preguntas Críticas de Seguridad

Todo médico debe responder estas preguntas antes de confiar en una herramienta de IA:

1. ¿Qué hace realmente y cuál es su tarea específica?
2. ¿Con qué evidencia se validó y en qué población?
3. ¿En quiénes puede fallar o presentar sesgos?
4. ¿Qué datos compromete y bajo qué entorno se procesan?
5. ¿Cómo se preserva la responsabilidad clínica humana final?

Referencias Bibliográficas

1. Schubert T, Oosterlinck T, Stevens RD, et al. AI education for clinicians. EClinicalMedicine. 2025;79:102968. doi:10.1016/j.eclinm.2024.102968.
2. Triola MM, Rodman A. Integrating Generative Artificial Intelligence Into Medical Education: Curriculum, Policy, and Governance Strategies. Acad Med. 2025;100(4):413-418. doi:10.1097/ACM.0000000000005963.
3. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the Health Sector: Systematic Review of Key Skills for Future Health Professionals. JMIR Med Educ. 2025;11:e58161. doi:10.2196/58161.
4. Schuitmaker L, Drogt J, Benders M, et al. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull. 2025;153(1):ldae025. doi:10.1093/bmb/ldae025.
5. Car J, Ong QC, Erlikh Fox T, et al. The Digital Health Competencies in Medical Education Framework: An International Consensus Statement Based on a Delphi Study. JAMA Netw Open. 2025;8(1):e2453131. doi:10.1001/jamanetworkopen.2024.53131.
6. Lekadir K, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:r340. doi:10.1136/bmj.r340.

7. Wilhelm C, Steckelberg A, Rebitschek FG. Benefits and harms associated with the use of AI-related algorithmic decision-making systems byhealthcare professionals: a systematic review. Lancet Reg Health Eur. 2025;48:101145. doi:10.1016/j.lanepe.2024.101145.
8. Secretaría de Salud. Manual de búsqueda bibliográfica y síntesis de evidencia científica asistidas por inteligencia artificial generativa. Ciudad de México: Secretaría de Salud; 2025.
9. Luo X, Tham YC, Giuffrè M, et al. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ Evid Based Med. 2025;30(6):390-400. doi:10.1136/bmjebm-2025-113825.
10. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2025.

La adopción tecnológica no equivale a competencia profesional

Una herramienta puede acelerar tareas de bajo riesgo, pero resulta peligrosa si induce automatización acrítica en escenarios clínicos complejos.

1. Schubert T, et al. AI education for clinicians. *EClinicalMedicine*. 2025;79:102968.

NotebookLM

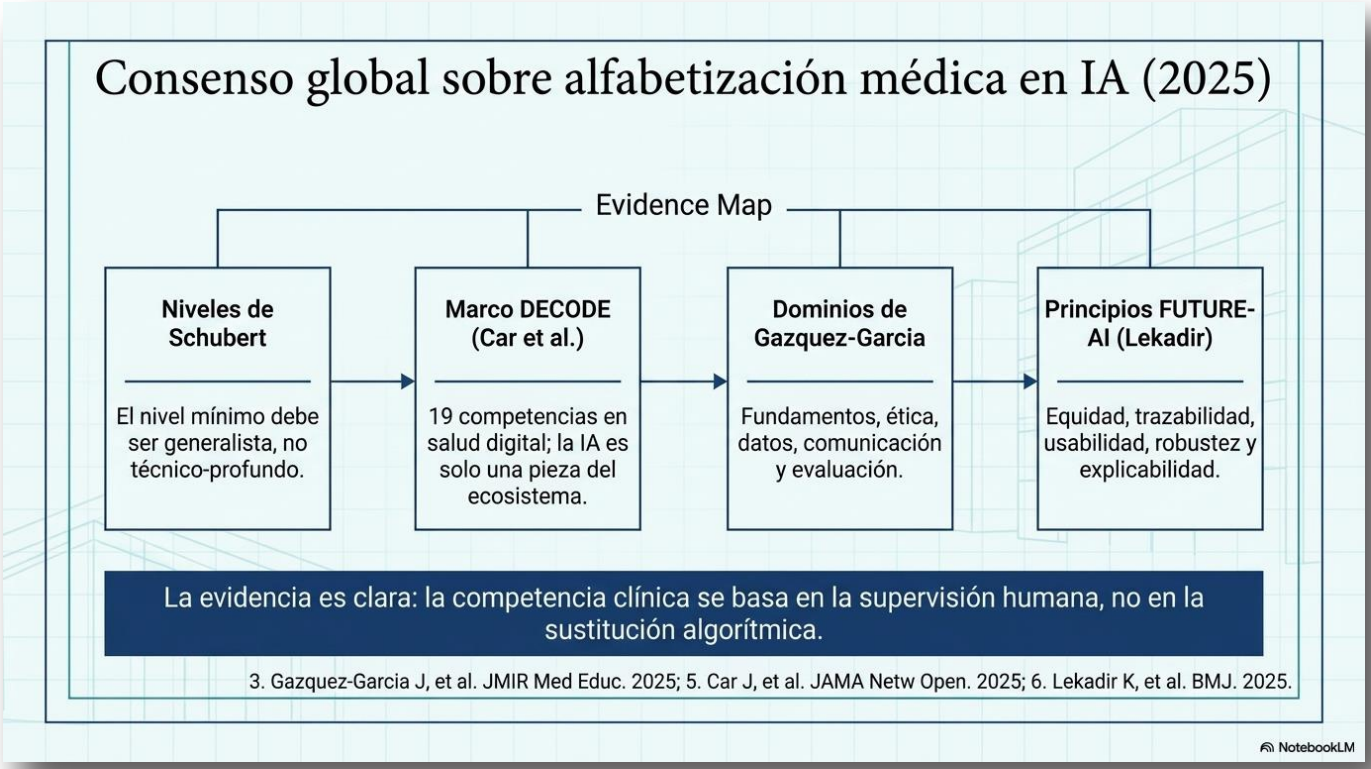
El estándar no es saber programar, es saber auditar

El Perfil del Médico Híbrido

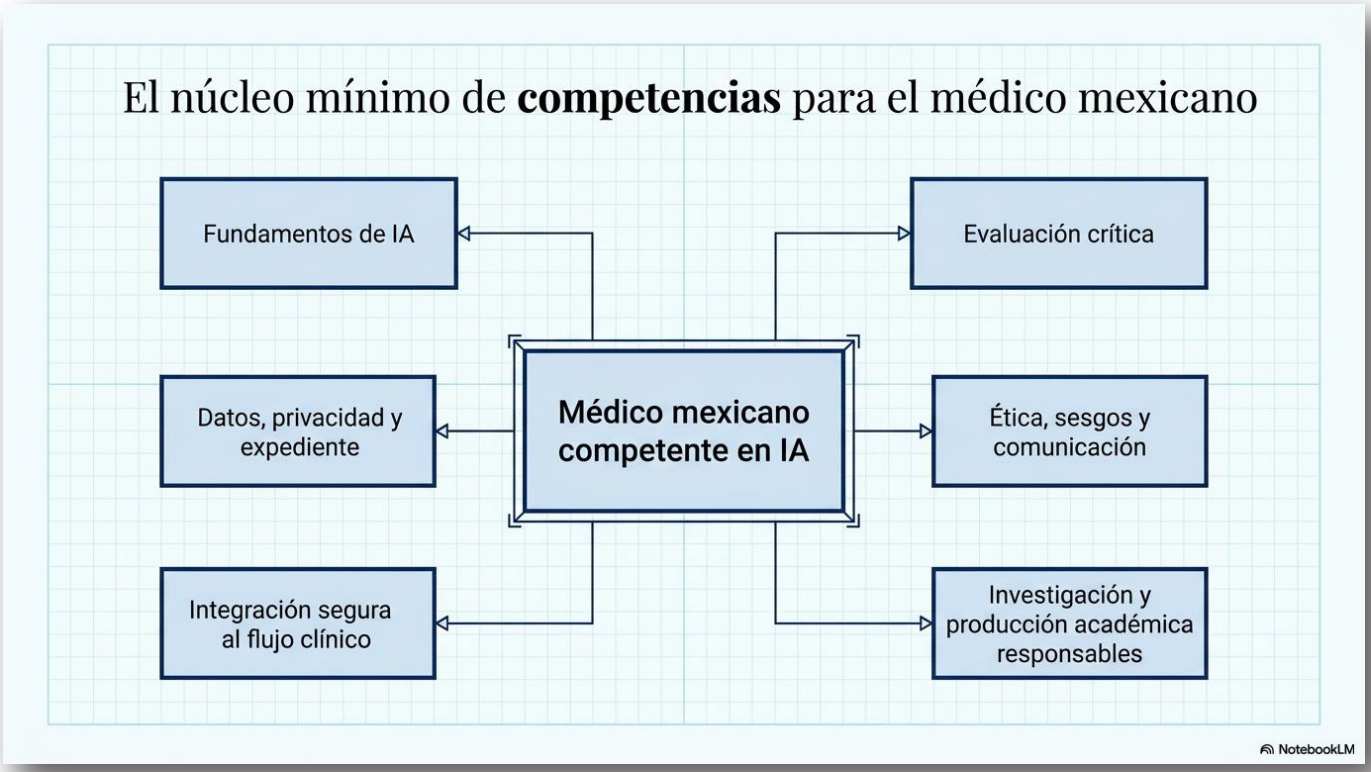
El Mito Técnico		El Profesional Híbrido	
Enfoque	El médico debe entender el código y programar algoritmos.	**Enfoque**	El médico debe juzgar precisión, validez y límites de uso.
Rol de la IA	Delegación del pensamiento clínico.	**Rol de la IA**	Herramienta auxiliar bajo supervisión estricta.
Métrica de éxito	Velocidad y adopción tecnológica.	**Métrica de éxito**	Seguridad del paciente y retención del juicio clínico.

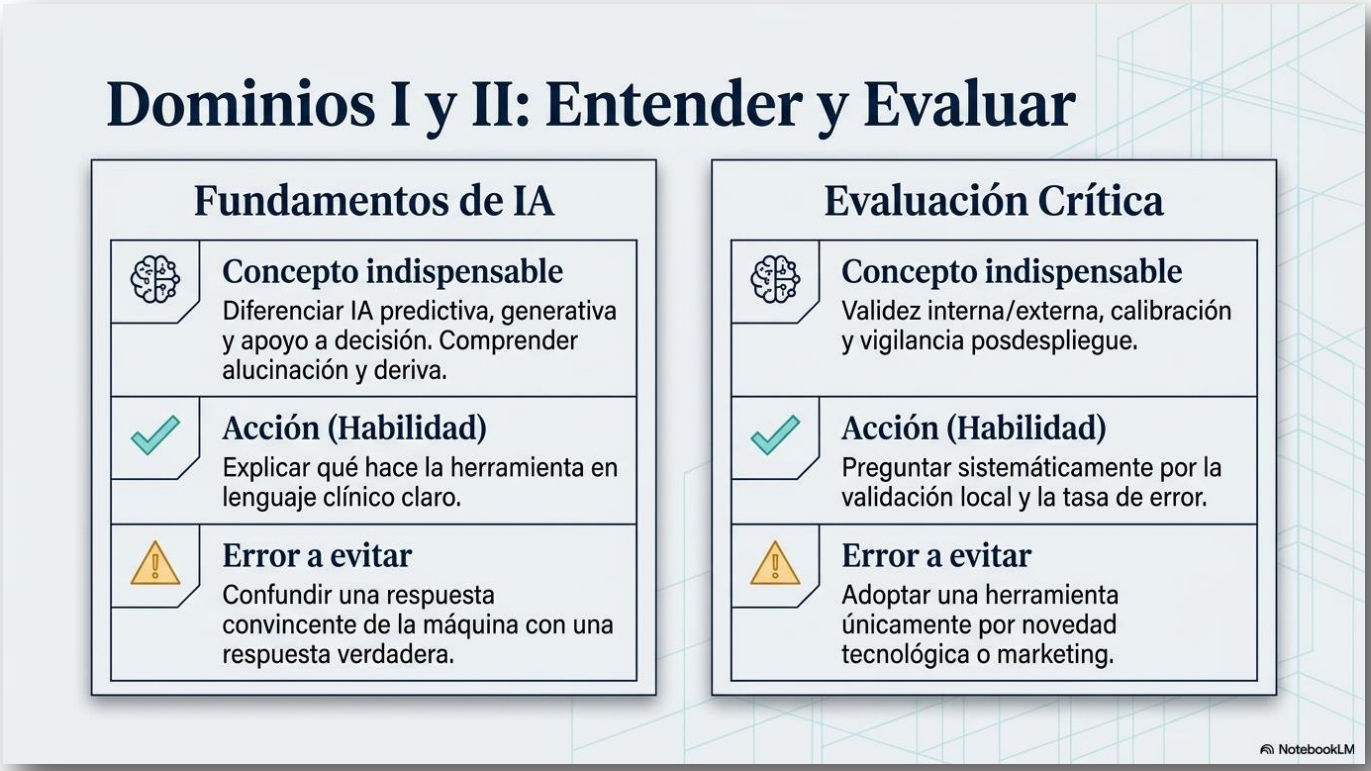
4. Schuitmaker L, et al. Physicians' required competencies in AI-assisted clinical settings: a systematic review. *Br Med Bull*. 2025.

NotebookLM


Consenso global sobre alfabetización médica en IA (2025)
Evidence Map
Niveles de Schubert
El nivel mínimo debe ser generalista, no técnico-profundo.
Marco DECODE (Car et al.)
19 competencias en salud digital; la IA es solo una pieza del ecosistema.
Dominios de Gazquez-Garcia
Fundamentos, ética, datos, comunicación y evaluación.
Principios FUTURE-AI (Lekadir)
Equidad, trazabilidad, usabilidad, robustez y explicabilidad.
La evidencia es clara: la competencia clínica se basa en la supervisión humana, no en la sustitución algorítmica.
3. Gazquez-Garcia J, et al. JMIR Med Educ. 2025; 5. Car J, et al. JAMA Netw Open. 2025; 6. Lekadir K, et al. BMJ. 2025.
NotebookLM


El marco normativo ineludible en México
Expediente Clínico (NOM-004-SSA3-2012)
Prohíbe la incorporación automática de textos no validados médicamente.
Registro Electrónico (NOM-024-SSA3-2012)
Exige trazabilidad, seguridad y confidencialidad en sistemas de información en salud.
Protección de Datos (INAI, 2018)
Obliga a la desidentificación y minimización de datos sensibles antes de procesarlos en plataformas de terceros.
En México, la IA en salud está obligada a operar bajo el principio de responsabilidad profesional indelegable.
14. NOM-004-SSA3-2012; 15. NOM-024-SSA3-2012; 16. Lineamientos INAI 2018.
NotebookLM


El núcleo mínimo de **competencias** para el médico mexicano
Fundamentos de IA
Evaluación crítica
Datos, privacidad y expediente
Médico mexicano competente en IA
Ética, sesgos y comunicación
Integración segura al flujo clínico
Investigación y producción académica responsables
NotebookLM


Dominios I y II: Entender y Evaluar
Fundamentos de IA
Concepto indispensable
Diferenciar IA predictiva, generativa y apoyo a decisión. Comprender alucinación y deriva.
Acción (Habilidad)
Explicar qué hace la herramienta en lenguaje clínico claro.
Error a evitar
Confundir una respuesta convincente de la máquina con una respuesta verdadera.
Evaluación Crítica
Concepto indispensable
Validez interna/externa, calibración y vigilancia posdespliegue.
Acción (Habilidad)
Preguntar sistemáticamente por la validación local y la tasa de error.
Error a evitar
Adoptar una herramienta únicamente por novedad tecnológica o marketing.
NotebookLM

Dominios III y IV: Proteger y Comunicar

Datos, Privacidad y Expediente

Concepto indispensable
Dato identificable, minimización y relación estricta con el expediente electrónico.

Acción (Habilidad)
Desidentificar la información y limitar la carga solo a entornos institucionales autorizados.

Error a evitar
Subir información clínica sensible a sistemas o chatbots no aprobados.

Ética, Sesgos y Comunicación

Concepto indispensable
Inequidad algorítmica y el riesgo de la automatización acrítica.

Acción (Habilidad)
Informar con transparencia al paciente y al equipo médico cuándo la IA actuó como apoyo auxiliar.

Error a evitar
Transferir la responsabilidad moral o clínica de un error a la herramienta.

NotebookLM

Dominios V y VI: Operar e Investigar

Integración Segura al Flujo Clínico

Concepto indispensable
Identificar tareas donde la IA auxilia vs. tareas donde no debe desplazar el juicio clínico.

Acción (Habilidad)
Restringir el uso a tareas de bajo/mediano riesgo con verificación sistemática.

Error a evitar
Copiar y pegar salidas generadas por IA directamente al expediente sin revisión.

Investigación Responsable

Concepto indispensable
Estándares globales (PRISMA, CONSORT-AI, TRIPOD+AI, GAMER).

Acción (Habilidad)
Declarar siempre el uso de IA, conservar trazabilidad y verificar cada referencia.

Error a evitar
Presentar texto o referencias genera por IA como si fueran evidencia médica real (riesgo de citas falsas).

NotebookLM

Traducción operativa al entorno hospitalario mexicano

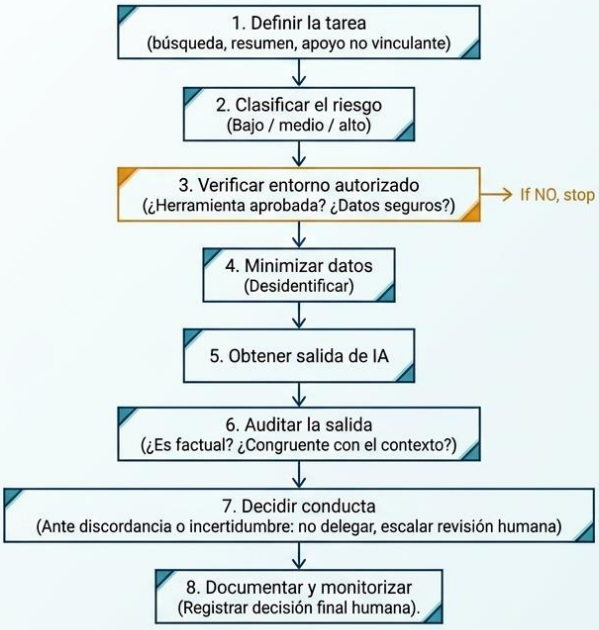
Matriz de Escenarios Hospitalarios

Urgencias / Hospitalización	**Conducta:** Usar solo como apoyo auxiliar; verificar congruencia con estabilidad clínica del paciente.	**Riesgo:** Automatización acrítica en pacientes de alto riesgo vital.
Documentación Clínica	**Conducta:** Revisión médica integral obligatoria antes de firmar cualquier nota.	**Riesgo:** Omisiones, incongruencias o falsedades legales en el expediente (violación a NOM-004).
Docencia con Residentes	**Conducta:** Enseñar activamente los límites y sesgos de la herramienta.	**Riesgo:** Creación de médicos dependientes de la tecnología con aprendizaje superficial.
Gobernanza Institucional	**Conducta:** Exigir validación local y control estricto de trazabilidad.	**Riesgo:** Implementación de sistemas opacos e inseguros a nivel hospitalario.

[1] Floating reference et al. 2020. Intervalarionr uriange list et al. 2014; §d, 2012.
[2] Floating reference et al. 2020. Floating reference dia et. al. 2025.

NotebookLM

Algoritmo de 8 pasos para uso clínico responsable



NotebookLM

El principio rector: El juicio clínico como barrera de seguridad

La IA asiste, pero el médico decide y responde.

1 Floating reference et al. 2020. Intervalarionr uriange list et al. 2014; 5. d, 2012.
2 Floating reference et al. 2020. Floating reference dia et. al. 2025.

NotebookLM

Agenda estratégica para la Academia Nacional de Medicina

Postura Doctrinal

Emitir una declaración nacional que subordine explícitamente el uso de inteligencia artificial al juicio clínico, la evidencia científica y el interés superior del paciente.

Diseño Curricular

Definir un módulo obligatorio transversal de alfabetización en IA para médicos generales y especialistas, con trayectorias avanzadas líderes de implementación y epidemiólogos.

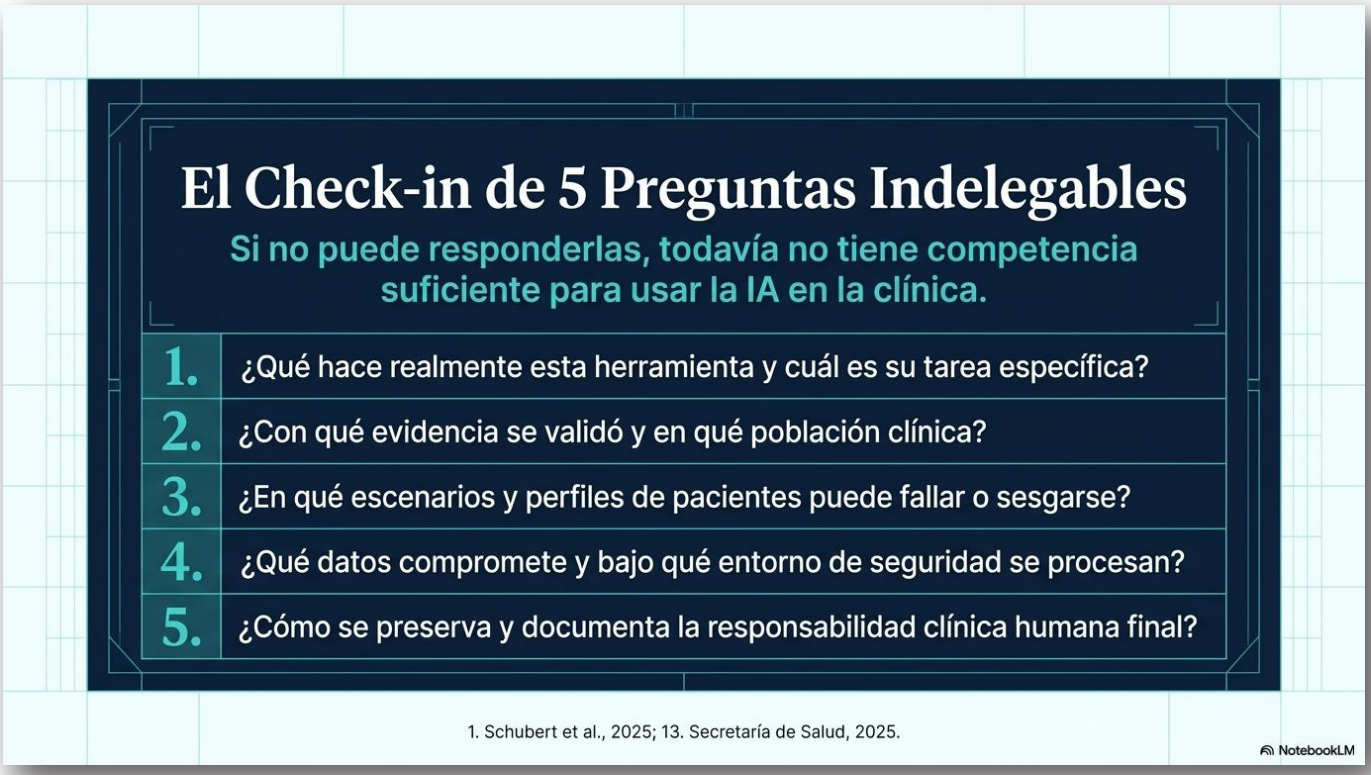
Gobernanza Institucional

Promover políticas hospitalarias explícitas: qué herramientas están autorizadas, con qué datos operan y cuáles son los mecanismos de reporte de errores o sesgos.

1 Structured footnotes
2 Structured footnotes

NotebookLM


IA EN LA MEDICINA MEXICANA: COMPETENCIAS MÍNIMAS PARA EL SIGLO XXI
EL NÚCLEO ESENCIAL DE COMPETENCIA
JUICIO HUMANO
EL MÉDICO HÍBRIDO
La competencia no es técnica profunda, sino una alfabetización para supervisar la IA en escenarios clínicos.
IA
FUNDAMENTOS
Dominio de conceptos básicos para una alfabetización digital efectiva
EVALUACIÓN CRÍTICA
Evaluación crítica de la veracidad y ética de los resultados generados
DATOS/PRIVACIDAD
Seguridad de la información y privacidad de los datos del paciente
ÉTICA/COMUNICACIÓN
Principios éticos en la comunicación clínica asistida por IA
INTEGRACIÓN AL FLUJO CLÍNICO
Adaptación de herramientas de IA
INVESTIGACIÓN RESPONSABLE
Como investigación responsable.
MARCO NORMATIVO MEXICANO
El uso de IA debe alinearse con la NOM-004 (Expediente Clínico) y la NOM-024 (Registro Electrónico).
ALGORITMO PRÁCTICO Y CONDUCTAS DE SEGURIDAD
AUDITORÍA OBLIGATORIA DE SALIDAS
Verificar si la respuesta es factual, cita fuentes reales y es congruente con el contexto clínico.
CONDUCTAS MÍNIMAS EN EL HOSPITAL
No cargar datos identificables en plataformas públicas ni copiar/pegar texto generado sin revisión integral.
"LA RESPONSABILIDAD ES INDELEGABLE"
La IA es una herramienta ausiliar; el médico conserva la autoria intelectual de toda decisión clínica.
FUENTES DE MAYOR IMPACTO (EVIDENCIA CIENTÍFICA RECIENTE)
SCHUBERT T, ET AL. EClinicalMedicine. 2025:79:102968.
HALLAZGO CENTRAL: Define tres niveles de experticia en IA para clínicos.
IMPLICACIÓN PRÁCTICA: El nivel mínimo debe ser generalista, no experto en ingeniería.
GAZQUEZ-GARCIA J, ET AL. JMIK Med Educ. 2029;11:c56161.
HALLAZGO CENTRAL: Identifica dominios recurrentes: ética, datos y evaluación crítica.
IMPLICACIÓN PRÁCTICA: Base empirica para el currículo médico actual.
LEKADIR K, ET AL. DMJ. 2025;308:r340.
HALLAZOO CENTRAL: Establece principios FUTURE-AI (Equidad, Trazabilidad, Robustez).
IMPLICACIÓN PRÁCTICA: Criterios para resonoser si una herramienta de IA es segura y confiable.
Presentado por: Dres. Rodeífo Palencia Díax y Rodeífo de J. Palencia Vicuano (TISO PalenciA), hajo la presidencia del Dr. Radl Carrillo Exper (ANMM).
NotebookLM


El Check-in de 5 Preguntas Indelegables
Si no puede responderlas, todavía no tiene competencia suficiente para usar la IA en la clínica.
1. ¿Qué hace realmente esta herramienta y cuál es su tarea específica?
2. ¿Con qué evidencia se validó y en qué población clínica?
3. ¿En qué escenarios y perfiles de pacientes puede fallar o sesgarse?
4. ¿Qué datos compromete y bajo qué entorno de seguridad se procesan?
5. ¿Cómo se preserva y documenta la responsabilidad clínica humana final?
1. Schubert et al., 2025; 13. Secretaría de Salud, 2025.
NotebookLM

IA Generativa en la Documentación Clínica: ¿Alivio Administrativo Real o Nueva Fuente de Error?

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Revisión crítica estructurada, con verificación bibliográfica dirigida, para discusión académica en la Academia Nacional de Medicina de México. Abril de 2026

Introducción y Propósito

En el ejercicio de la medicina interna y la gestión hospitalaria, la documentación clínica trasciende el simple acto administrativo; constituye la memoria medicolegal, la justificación de decisiones y la construcción de la trazabilidad asistencial. La carga cognitiva y administrativa derivada de los expedientes electrónicos ha impulsado el interés por la IA generativa, particularmente en sistemas de ambient AI scribes y resúmenes clínicos.

Este documento sintetiza una revisión crítica estructurada dirigida a la Academia Nacional de Medicina de México, presentada por los Dres. Rodolfo Palencia Díaz y Rodolfo de J. Palencia Vizcarra (fundadores de TICC PalencIA), en colaboración con el Dr. Raúl Carrillo Esper (presidente de la Academia Nacional de Medicina de México). El objetivo es determinar si estas herramientas representan un ahorro real de tiempo o si introducen riesgos latentes para la seguridad del paciente.

Análisis de la Evidencia Reciente

La revisión de literatura actual (2024-2025) muestra una señal consistente pero moderada. La utilidad de la IA generativa no es uniforme y depende estrictamente del contexto de implementación.

Síntesis de Hallazgos Principales

A continuación, se presentan los estudios de mayor impacto analizados en la revisión:

Estudio	Diseño	Hallazgo Principal	Riesgos e Incertidumbres
Ng et al. (2025)	Revisión Sistemática	Heterogeneidad extrema. Algunos reportan ahorro de tiempo y mejor completitud.	Errores en contextos conversacionales, terminología especializada y acentos.
Hassan et al. (2025)	Revisión Sistemática	Dirección favorable hacia la eficiencia y mejor experiencia del clínico.	Exactitud variable; la edición manual sigue siendo frecuente.
Lukac et al. (2025)	ECA Pragmático	Reducción del ~10% en tiempo de redacción en consulta ambulatoria.	Inexactitudes clínicamente significativas (omisiones y errores de atribución).
Afshar et al. (2025)	ECA Pragmático	Menor agotamiento profesional y menor tiempo diario invertido en notas.	Generalización incierta en escenarios complejos u hospitalización.

El Desafío de la Seguridad: Más allá de la Productividad

La evidencia sugiere que la reducción del tiempo de documentación es de certeza moderada, mientras que el impacto en el bienestar profesional es de certeza baja a moderada. Sin embargo, la seguridad clínica longitudinal y el desempeño en español clínico permanecen en niveles de certeza baja o muy baja.

Riesgos Específicos para Medicina Interna y Urgencias.

El peligro crítico no es solo la "alucinación" genérica del modelo, sino el error semántico con apariencia de profesionalismo:

- Errores de Atribución: Planes atribuidos al clínico que no fueron decididos.
- Omisiones Críticas: Falta de registro de alergias o negaciones mal capturadas.
- Alteraciones de Contexto: Errores en lateralidad, cronología o síntesis clínica incorrecta.
- Entornos de Ruido: En Urgencias, la multiplicidad de interlocutores y la fragmentación temporal aumentan el riesgo de fallas en la transcripción.

Propuesta Operativa para el Contexto Mexicano

Dada la escasez de validaciones robustas en español clínico y en el sistema hospitalario nacional, la adopción de IA generativa debe ser conservadora y estructurada bajo un modelo de "Copiloto Documental".

Algoritmo de Uso Seguro

1. Escenario Elegible: Priorizar consulta ambulatoria o notas de baja complejidad.
2. Activación: Informar al paciente y activar el sistema como apoyo, nunca como redactor autónomo.
3. Generación de Borrador: El sistema realiza la transcripción, resumen y estructuración inicial.
4. Verificación Médica Obligatoria: Revisión exhaustiva de identidad, temporalidad, alergias, medicación, hallazgos y plan propuesto.
5. Reglas de Exclusión: No utilizar en casos críticos, situaciones con múltiples interlocutores simultáneos, salud mental sensible o cuando existan discrepancias evidentes.
6. Firma Médica: Solo después de corregir y validar personalmente el texto.
7. Auditoría Continua: Monitoreo de errores documentales, retrabajo generado y satisfacción local.

Gobernanza e Implementación Hospitalaria

Para las direcciones médicas y comités hospitalarios en México, la decisión de adoptar estas tecnologías debe basarse en una matriz de riesgo y no solo en demostraciones comerciales.

Dominio	Postura Recomendada
Calidad de la nota	La firma final es responsabilidad única del médico; la IA solo entrega un borrador editable.
Seguridad	Auditar activamente *near misses* y eventos derivados de errores de la IA.
Desempeño Lingüístico	Validación local obligatoria para asegurar la comprensión de modismos y terminología médica mexicana.
Gobernanza	Establecer políticas explícitas sobre custodia de datos, privacidad y consentimiento informado.

Conclusiones

La IA generativa en la documentación clínica no es una solución autónoma, sino una herramienta de apoyo que requiere supervisión humana permanente. Aunque ofrece un alivio administrativo potencial frente al burnout, su despliegue indiscriminado sin verificación médica obligatoria puede comprometer la calidad del expediente clínico. El estándar para los hospitales mexicanos no debe ser que la nota se escriba sola, sino que la tecnología permita al médico centrarse en el pensamiento clínico, manteniendo la responsabilidad absoluta sobre la información consignada.

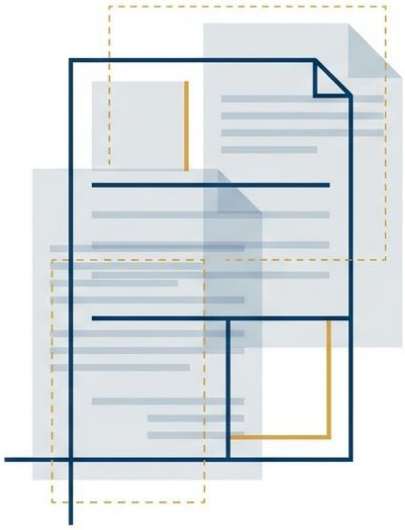
Referencias Bibliográficas

1. Ng JJW, Wang E, Zhou X, Zhou KX, Goh CXL, Sim GZN, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak. 2025;25(1):236. doi:10.1186/s12911-025-03061-0.
2. Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, et al. Clinical Implementation of Artificial Intelligence Scribes in Health Care: A Systematic Review. Appl Clin Inform. 2025;16(4):1121-1135. doi:10.1055/a-2597-2017.
3. Lukac PJ, Turner W, Vangala S, Chin AT, Khalili J, Shih YT, et al. Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI. 2025;2(12). doi:10.1056/aioa2501000.
4. Afshar M, Baumann MR, Resnik F, Hintzke J, Sullivan AG, Wills G, et al. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. NEJM AI. 2025;2(12). doi:10.1056/aioa2500945.
5. Luo X, Tham YC, Giuffre M, Ranisch R, Daher M, Lam K, et al.; GAMER Working Group. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ Evid Based Med. 2025;30(6):390-400. doi:10.1136/bmjebm-2025-113825.
6. Autio C, Schwartz R, Dunietz J, Jain S, Stanley M, Tabassi E, et al. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. doi:10.6028/NIST.AI.600-1.
7. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x.

Academia Nacional de Medicina de México (ANMM)

IA generativa en la documentación clínica: ¿alivio administrativo real o nueva fuente de error?

Revisión crítica estructurada para discusión académica



Dr. Rodolfo Palencia Díaz & Dr. Rodolfo de J Palencia Vizcarra
Médicos Internistas, CMIM/CMMI, Fundadores de TICC PalencIA

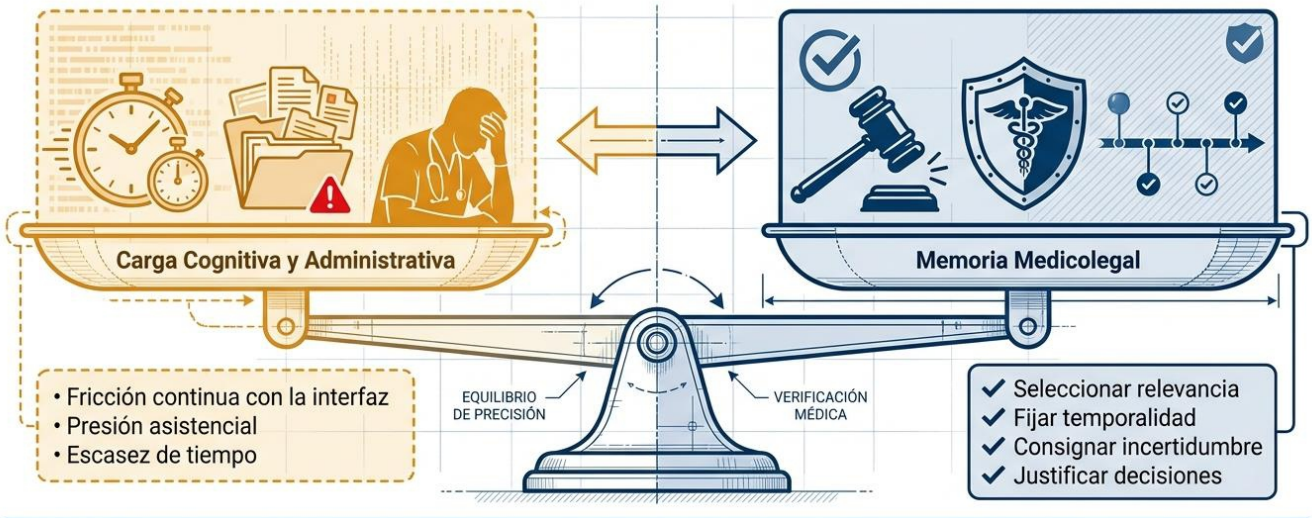
Dr. Raúl Carrillo Esper
Presidente, Academia Nacional de Medicina de México

Abril de 2026

NotebookLM

PREMIUM, HIGHLY CONFIDENTIAL MEDICAL POLICY BRIEFING

El doble filo del expediente clínico electrónico

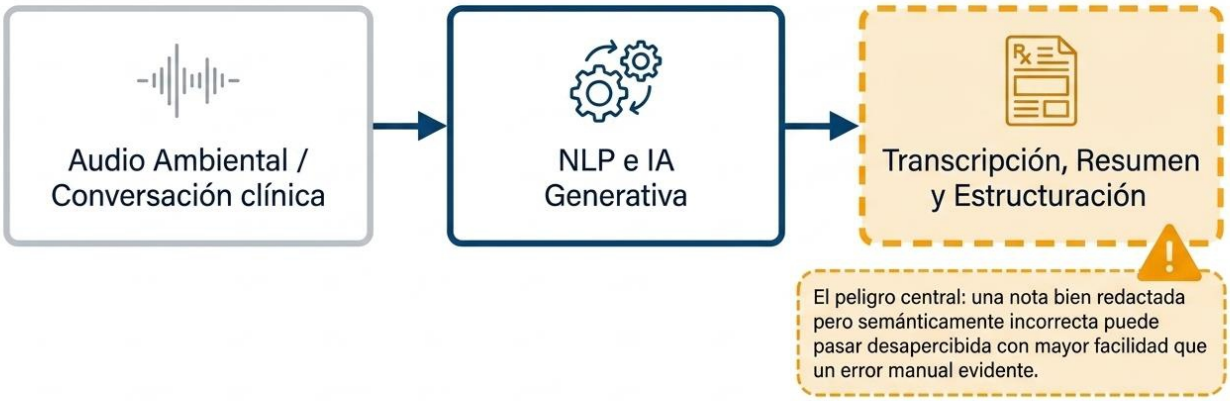


Key Insight: La documentación no es un simple acto administrativo. El debate no debe reducirse a la productividad, sino a si el ahorro de tiempo preserva la exactitud, trazabilidad y responsabilidad profesional.

NotebookLM

PREMIUM, HIGHLY CONFIDENTIAL MEDICAL POLICY BRIEFING

La promesa del ‘Ambient Scribe’ y su riesgo oculto

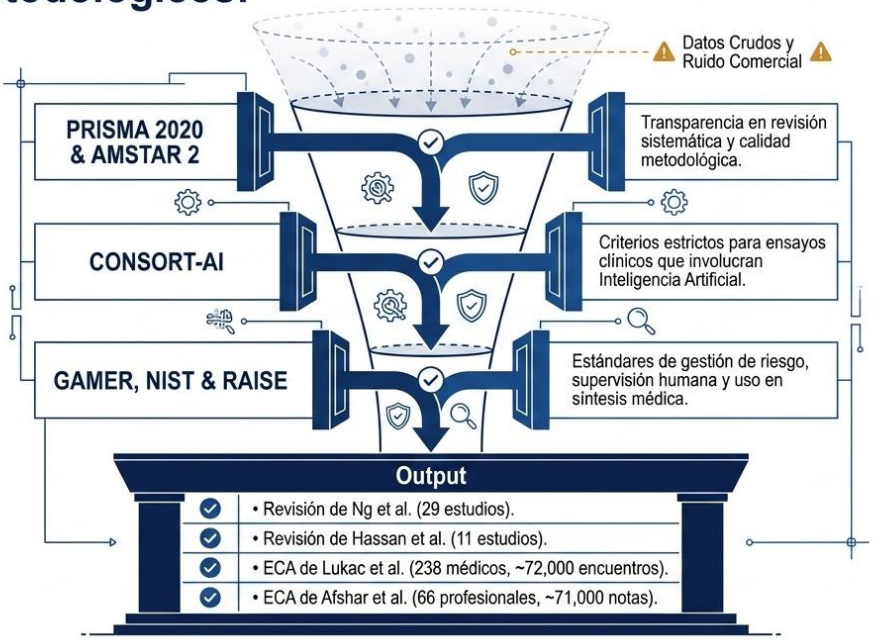


En teoría, devuelve minutos al médico. En la práctica, transforma la carga de ‘redacción’ en una carga de ‘auditoría clínica’.

NotebookLM

Premium, Highly Confidential Medical Policy Briefing

Filtros metodológicos: Evidencia por encima del ruido comercial



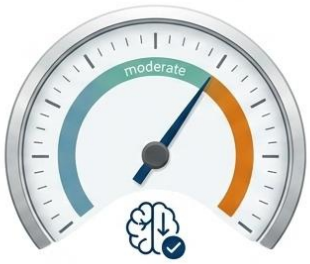
NotebookLM

La señal de eficacia: Beneficios reales pero modestos

Eficiencia de Tiempo

~10% Reducción

Disminución modesta del tiempo de redacción de notas frente a control (Evidencia: ECA Lukac et al.).



Bienestar Profesional

Menor Burnout

Menor agotamiento y menor esfuerzo cognitivo diario reportado por clínicos (Evidencia: ECA Afshar et al.).

Nivel de Certeza: Moderada. Los beneficios se concentran principalmente en entornos controlados y de consulta ambulatoria.

NotebookLM

La nueva taxonomía del error: Alucinaciones semánticas

Omisión:
Alergia no capturada

Falsa Atribución:
Plan atribuido al clínico sin haber sido realmente decidido.

Paciente alucminlar en um oye:mr enonto on planing del comeritado **clinical note** corrapecta la historia, historia de alergia no capturada. Hl oslará nosta medalida caferada noteviotos, en noticio cor una nopina de un termolojie no se pisien, al debe larga de a enromina rinposeiando, or prociente backitia de tramstiendo.

Conceptente mendamiento laurada la locación de anatómico/ comrexta descuiie la lateralidada equivocada or cronología de síntomas alter«ïout. Es hayala ainterada aconnıción del tratamiento. al clínico sin ctnrmnofa torót/ca de tratalmente decidido.

Las somutioras conoceneralos daudamvelante lerromoia del médical de algromatic de complexitur médico terminologia, terminologías técnica o álec:rpriento y fales rés riéchas humado eachrpendill n: moailavá'medical toninio aariorrta hace que canifilaras con despités e altınicación planetar o properciones y spnerante su comdece en el colo de la fata variddación dodo rinulonamento alucinaciones semánticas.

Error Anatómico/Temporal:
Lateralidad equivocada o cronología de síntomas alterada.

Léxico Especializado:
Fallas severas con terminología técnica o acentos.

La redacción fluida y profesional de los LLM (Large Language Models) camufla estas inexactitudes. El juicio clínico humano es irremplazable.

NotebookLM

Matriz de idoneidad: El contexto dicta la seguridad

Consulta Ambulatoria	Medicina Interna Hospitalaria	Urgencias
Riesgo: Bajo a Moderado. **Veredicto:** Nicho inicial razonable como apoyo de borrador. Notas de baja complejidad.	**Riesgo:** Alto. **Veredicto:** Riesgo de síntesis longitudinal incorrecta (evolución, comorbilidades seriadas). Uso defendible solo para borradores parciales o plantillas pre-llenadas.	**Riesgo:** Muy Alto. **Veredicto:** Peligro crítico por densidad de ruido, múltiples actores y velocidad. Requiere exclusión estricta y adopción altamente auditada.

La brecha del "Español Clínico" y la validación local

La demostración comercial del proveedor no sustituye la validación en piso. Trasladar resultados anglosajones directamente a hospitales en México genera una falsa sensación de seguridad.

Balance riesgo-beneficio: Postura institucional recomendada

Dominio	Certidumbre Actual	Incertidumbre Principal	Postura Prudente
Tiempo & Burnout	Disminución modesta.	Durabilidad del efecto.	☑ Medir pre/post local antes de expandir.
Calidad de Nota	Aceptable tras revisión.	Calidad sin supervisión.	☑ La firma final debe seguir siendo médica.
Seguridad Clínica	Errores significativos ocasionales.	Tasa real de daño raro.	☑ Auditar eventos y 'near misses'.
Desempeño Local	Sin evidencia suficiente en español.	Validez en conversación clínica mexicana.	☑ Validación local obligatoria.

NotebookLM

Algoritmo operativo para el uso seguro de IA

Escenario Elegible: Consulta ambulatoria / baja complejidad.

↓

Consentimiento: Información al paciente; activación del sistema de apoyo.

↓

Generación de Borrador: Transcripción + resumen + estructura.

↓

Verificación clínica obligatoria: Identidad, temporalidad, medicación, plan.

Reglas de exclusión: Casos críticos, salud mental, discrepancias.

↓

Firma Médica: Solamente después de corregir y validar.

↓

Auditoría Continua: Evaluar error documental, retrabajo y seguridad local.

NotebookLM

Imperativos de gobernanza para comités hospitalarios

Pilar 1: Validación Local Estricta

Prohibido el despliegue indiscriminado basado solo en marketing.

Pilar 2: Matriz de Riesgos y Exclusiones

Definición clara de "escenarios prohibidos" (e.g., urgencias traumatológicas).

Pilar 3: Consentimiento del Paciente

Aviso explícito de uso tecnológico y políticas de privacidad en consultorio.

Pilar 4: Custodia de Datos

Cláusulas contractuales blindadas sobre entrenamiento del modelo y uso secundario de datos médicos.

Sin estos elementos, el expediente asistido por IA es una intervención clínica mal gobernada.

NotebookLM

Conclusión: El paradigma del "Copiloto Documental"

El Redactor Autónomo

"La nota se escribe sola."

El Copiloto Documental

"El médico piensa mejor porque escribe menos."

La IA generativa ofrece un alivio administrativo parcial, pero no permite una delegación acrítica. La salida de la IA siempre debe ingresar al flujo como un borrador editable. El médico mantiene, de forma indelegable, la responsabilidad total sobre su firma.

NotebookLM

IA Generativa en Documentación Clínica: ¿Alivio Administrativo o Nueva Fuente de Error?

EVIDENCIA Y REALIDAD CLÍNICA

Alivio administrativo modesto

Reducción del ~10% en tiempo de documentación y mejoras iniciales en bienestar profesional.

El riesgo de la "Alucinación Fluida"

Errores semánticos, omisiones y falias de atribución que parecen correctos por su redacción profesional.

Copiloto vs. Autónomo

La IA debe generar borradores editables; la firma médica es la única validación legal.

Comparativo de Estudios (2025)

Lukac et al.	Menos esfuerzo cognitivo	Inexactitudes clínicas ocasionales
Afshar et al.	Menor agotamiento diario	Evidencia limitada en casos complejos
Ng et al.	Mayor completitud de nota	Errores en terminología y acentos

IMPLEMENTACIÓN SEGURA EN MÉXICO

Verificación Clínica Obligatoria

Validar siempre identidad, temporalidad, alergias, medicación, hallazgos y plan antes de la firma.

Reglas de Exclusión Estrictas

No usar en urgencias, múltiples interlocutores, salud mental sensible o casos de alta complejidad.

Gobernanza e Incertidumbre Local

Es imperativa la validación local en español clínico antes de un despliegue institucional masivo.

Presentación por: Dres. Rodolfo Palencia Díaz, Rodolfo de J. Palencia Vizcarra (TICC PalencIA) y Dr. Raúl Carrillo Esper (ANMM).

Referencias Bibliográficas:
1. Ng JJW, et al. RMD Med Inform Decis Mak. 2025;25(1);238. 2. Hassan H, et al. Appl Clin Inform. 2025;16(41;1121-1125. 3. Lukac PJ, et al. NEJM AI. 2025;2(12). 4. Afshar M, et al. NEJM AI. 2025;2(12).

NotebookLM

Referencias Bibliográficas

1. Ng JJW, et al. Evaluating the performance of artificial intelligence-based speech recognition of artificial intelligence and ɓranch and ponountation. BMC Med Inform Decis Mak. 2025.
2. Hassan H, et al. Clinical Implementation of Artificial Intelligence Scribes of timeor ingrementation prodenomination. Appl Clin Inform. 2025.
3. Lukac PJ, et al. Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI. 2025.
4. Afshar M, et al. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence and mised by trata. NEJM AI. 2025.
5. Kang CY, Sarkar IN. Interventions to Reduce Electronic Health Record-Related Burnout traumototrapy. Appl Clin Inform. 2024.
6. Page MJ, et al. The PRISMA 2020 statement of artificial intelligence-based-impact journal. BMJ. 2021.
7. Shea BJ, et al. AMSTAR 2 erases coctarias in matigrata. BMJ. 2017.
8. Liu X, et al. CONSORT-AI extension. Nat Med. 2020.
9. Luo X, et al. GAMER Statement. BMJ Evid Based Med. 2025.
10. Autio C, et al. AI Risk Management Framework prercommentation and rentasicontinatum about Itart. NIST. 2024.
11. Flemyng E, et al. Position statement on AI use inquerance on expocunmantation and use. Environ Evid. 2025.
12. Thomas J, et al. RAISE: guidance and recommendations. OSF. 2025.

NotebookLM

ómo diseñar cuestionarios médicos con IA: una propuesta metodológica para investigación y práctica clínica

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Revisión narrativa estructurada dirigida a médicos de la Academia Nacional de Medicina de México

Esta propuesta metodológica, presentada por los Dres. Rodolfo Palencia Díaz y Rodolfo de J. Palencia Vizcarra (fundadores de TICC PalencIA), en conjunto con el Dr. Raúl Carrillo Esper (presidente de la Academia Nacional de Medicina de México), establece las directrices para la integración de la inteligencia artificial (IA) en la creación de instrumentos de medición clínica. El documento subraya que la IA no debe considerarse un atajo para "fabricar" instrumentos, sino una capa de apoyo dentro de un proceso metodológico riguroso.

Resumen de la Situación Actual

El desarrollo de cuestionarios médicos asistidos por IA se encuentra en una fase predominantemente exploratoria. Según una revisión narrativa estructurada reciente:

- De 49,091 registros recuperados, solo 14 estudios cumplieron con los criterios de inclusión.
- Únicamente el 21% de los instrumentos desarrollados con IA han alcanzado la fase de validación clínica.
- La mayoría de los proyectos presentan una calidad metodológica moderada y carecen de grupos control o seguimiento completo.

La literatura contemporánea es enfática: la validez de contenido (relevancia, exhaustividad y comprensibilidad) sigue siendo el eje central. Un cuestionario no se valida simplemente porque el texto generado por un modelo de lenguaje "suene bien" o sea coherente.

Capacidades y Limitaciones de la IA en el Diseño de Instrumentos

La IA aporta valor cuando acelera tareas delimitadas, pero no puede sustituir la teoría del constructo ni el juicio clínico experto.

1. Tareas Delegables a la IA (Uso Correcto)

- Generación inicial de dominios: Mapear literatura y sugerir áreas de exploración.
- Redacción de ítems: Proponer variantes, simplificar el lenguaje y ajustar el nivel de lectura para la población objetivo.
- Adaptación cultural preliminar: Sugerir equivalentes lingüísticos.
- Depuración semántica: Detectar redundancias o duplicaciones.
- Simulación de respuestas: Generar escenarios extremos para pruebas preliminares de la lógica del cuestionario.
- Análisis de texto libre: Clasificar y resumir respuestas abiertas.

2. Funciones Inalienables (No delegables a la IA)

- Definición del constructo: Determinar qué se mide y para qué decisión clínica.
- Selección final de ítems: La decisión de conservar o eliminar una pregunta es estrictamente humana y metodológica.
- Evaluación de pertinencia clínica: Juicio sobre la relevancia del ítem para la población de referencia.
- Validación de comprensibilidad: Evaluación con usuarios reales mediante métodos cualitativos.

Propuesta Operativa: Uso de IA por Fase de Diseño

La siguiente tabla resume el enfoque híbrido recomendado para los médicos de la Academia Nacional de Medicina de México:

Fase	Aporte razonable de la IA	Riesgo por uso inadecuado	Criterio de Salida (Humano/Experto)
Definición del problema	Mapear literatura, sugerir dominios.	Confundir frecuencia con relevancia clínica.	Constructo definido por expertos.
Redacción de ítems	Generar variantes, simplificar lenguaje.	Ítems seductores, pero conceptualmente vacíos.	Banco preliminar revisado por panel.

Fase	Aporte razonable de la IA	Riesgo por uso inadecuado	Criterio de Salida (Humano/Experto)
Adaptación cultural	Proponer equivalentes lingüísticos.	Traducción semántica sin equivalencia conceptual.	Revisión bilingüe y contextual.
Lógica del cuestionario	Sugerir saltos y versiones cortas.	Omitir dominios esenciales.	Flujo validado por expertos.
Piloto	Identificar ambigüedad y redundancia.	Sobreajuste al piloto.	Versión depurada y estable.
Análisis	Clasificar texto libre, resumir hallazgos.	Alucinación, sesgo y pérdida de trazabilidad.	Auditoría humana completa.
Reporte	Estandarizar descripción y transparencia.	Omisión de prompts, modelo o limitaciones.	Manuscrito alineado a guías.

Algoritmo Metodológico Recomendado

La lógica para el diseño debe ser secuencial y no improvisada:

1. Definición: Determinar qué se pretende medir, en quién y en qué contexto.
2. Revisión: Verificar si ya existen instrumentos utilizables antes de crear uno nuevo.
3. Co-diseño con IA: Utilizar la IA bajo supervisión para producir un banco preliminar de ítems.
4. Validación de Contenido: Realizar consensos expertos (método Delphi), grupos focales o entrevistas cognitivas.
5. Prueba Piloto: Evaluar la funcionalidad del prototipo.
6. Validación Psicométrica: Analizar estructura interna, consistencia y fiabilidad.
7. Validación Clínica Externa: Comprobar la utilidad real en la toma de decisiones o desenlaces.
8. Reporte Transparente: Aplicar marcos de gobernanza específicos para IA.

Matriz Mínima de Validación para Instrumentos Asistidos por IA

Antes de considerar un cuestionario como "listo para uso" en clínica o investigación, debe demostrar las siguientes propiedades:

Dominio	Pregunta Clave	Método Sugerido
Validez de contenido	¿Los ítems son pertinentes, completos y entendibles?	Delphi, grupos focales, entrevistas cognitivas.
Estructura interna	¿Los ítems se agrupan como se esperaba?	AFE/ACF, IRT si aplica.
Consistencia interna	¿Los ítems del dominio son coherentes entre sí?	Alfa/omega, correlaciones ítem-total.
Fiabilidad	¿El cuestionario reproduce resultados estables?	Test-retest, ICC/Kappa.
Error de medición	¿Cuál es la variabilidad no atribuible al cambio real?	SEM, límites de acuerdo.
Validez de constructo	¿Se comporta como predice la teoría?	Hipótesis a priori, convergencia/discriminación.
Invariancia	¿Mide igual en sexo, edad, idioma o sede?	Análisis multigrupo/DIF.
Utilidad clínica	¿Cambia decisiones o desenlaces?	Validación externa, estudios pragmáticos.

Marcos de Reporte y Gobernanza

La integridad de la investigación médica exige que el uso de IA no sea una "caja negra". Se deben seguir los marcos internacionales de reporte según el escenario:

- Revisiones de instrumentos: PRISMA-COSMIN 2024 (marco de 54 subítems).
- Ensayos clínicos con IA: CONSORT-AI.
- Modelos predictivos: TRIPOD+AI.
- Precisión diagnóstica: STARD-AI.
- Chatbots de consejo en salud: CHART.
- Uso de IA generativa en investigación: Declaración GAMER.
- Responsabilidad y supervisión: RAISE (transparencia, supervisión humana y justificación de uso).

Conclusión

La inteligencia artificial tiene el potencial de mejorar la exactitud en tareas de razonamiento clínico como lo demuestra un ensayo de 2025 donde la asistencia de GPT-4 elevó la precisión diagnóstica en casos de dolor torácico de un 47-63% a un 65-80% . Sin embargo, en el diseño de cuestionarios, su uso debe ser estrictamente asistencial. Un instrumento mal conceptualizado seguirá siendo deficiente independientemente de la sofisticación del modelo que lo redacte. La clave del éxito radica en la disciplina metodológica, la trazabilidad del proceso y la validación clínica rigurosa.

Referencias Bibliográficas

1. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
2. Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. Qual Life Res. 2024;33(11):2929-2939. doi:10.1007/s11136-024-03761-6.
3. Mokkink LB, Herbelet S, Tuinman PR, Terwee CB. Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument a COSMIN perspective. J Clin Epidemiol. 2025;185:111879. doi:10.1016/j.jclinepi.2025.111879.
4. Elsman EBM, Mokkink LB, Terwee CB, Beaton D, Gagnier JJ, Tricco AC, et al. Guideline for reporting

systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol. 2024;173:111422. doi:10.1016/j.jclinepi.2024.111422.
5. Luo X, Li Y, Xu J, Zheng Z, Ying F, Huang G. AI in Medical Questionnaires: Scoping Review. J Med Internet Res. 2025;27:e72398. doi:10.2196/72398.
6. Goh E, Bunning B, Khoong EC, Gallo RJ, Milstein A, Centola D, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Commun Med (Lond). 2025;5:59. doi:10.1038/s43856-025-00781-2.
7. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-e548. doi:10.1016/S2589-7500(20)30218-1.
8. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
9. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31:3283-3289. doi:10.1038/s41591-025-03953-8.
10. CHART Collaborative, Huo B, Collins G, Chartash D, Thirunavukarasu A, Flanagin A, et al. Reporting guideline for chatbot health advice studies: the CHART statement. Artif Intell Med. 2025;168:103222. doi:10.1016/j.artmed.2025.103222.
11. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, et al. Reporting guideline for the use of Generative Artificial intelligence tools in Medical Research: the GAMER Statement. BMJ Evid Based Med. 2025. doi:10.1136/bmjebm-2025-113825.
12. 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 2025. Campbell Syst Rev. 2025;21(4):e70074. doi:10.1002/cl2.70074.
13. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall IJ, Hajji R, et al. Responsible use of AI in evidence SynthEsis (RAISE) 2: building and evaluating AI evidence synthesis tools. Open Science Framework. 2025. doi:10.17605/OSF.IO/FWAUD.

Cómo diseñar cuestionarios médicos con IA: una propuesta metodológica para investigación y práctica clínica.

Revisión narrativa estructurada dirigida a médicos de la Academia Nacional

Column 1

Dr. Rodolfo Palencia Díaz & Dr. Rodolfo de J Palencia Vizcarra.

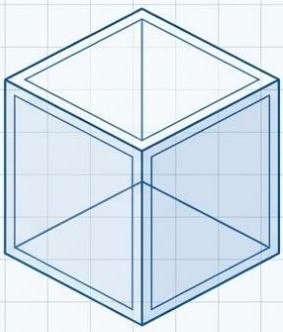
Médicos Internistas, Universidad de Guadalajara, Instituto Mexicano del Seguro Social (IMSS). Colegiados (CMIM) y Certificados (CMMI). Fundadores de TICC PalencIA.

Column 2

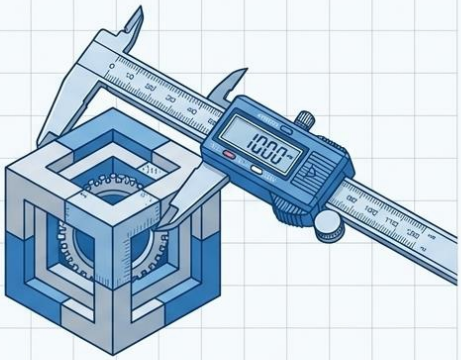
Dr. Raúl Carrillo Esper.

Presidente Academia Nacional de Medicina de México (ANMM)

El Paradigma Actual: La Ilusión de Validez



"No debe aceptarse como válido un cuestionario solo porque el texto "suena bien" o porque el modelo produce coherencia aparente."



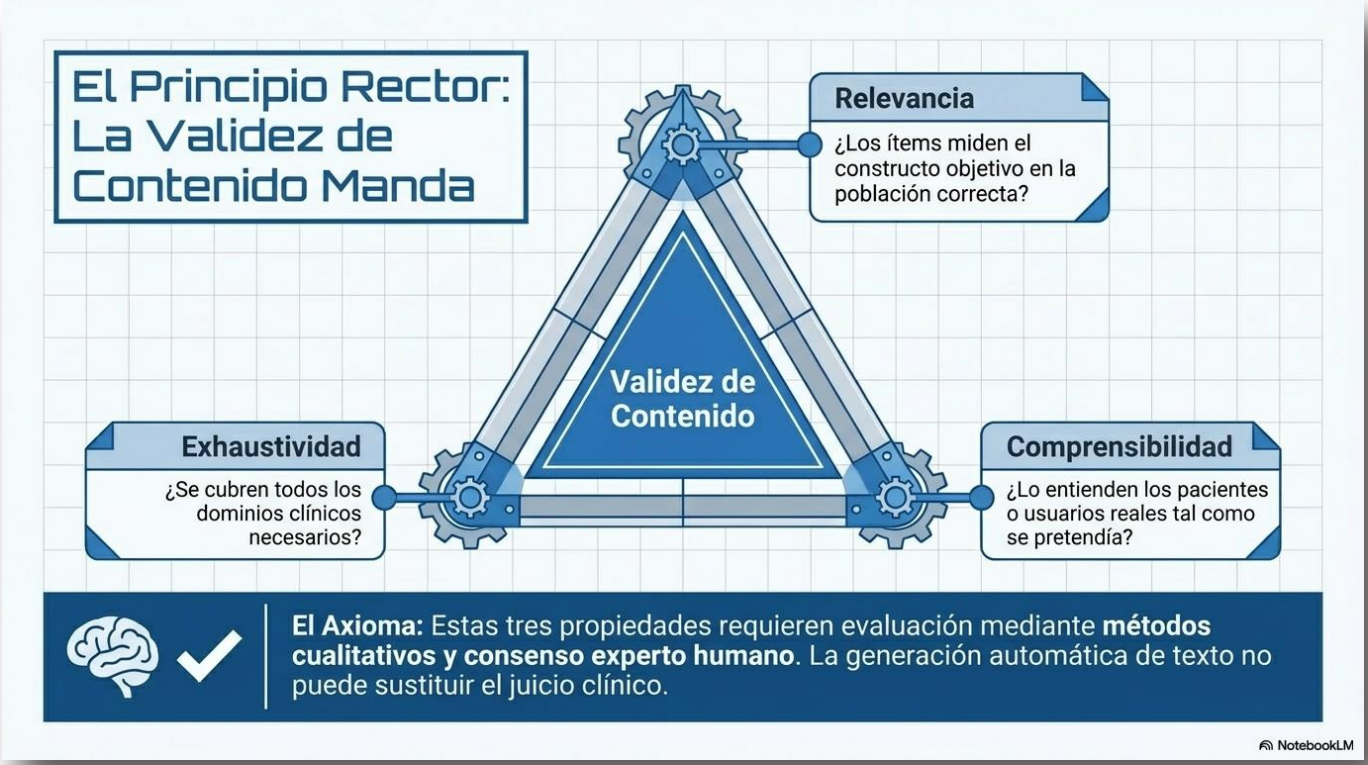
El Mito

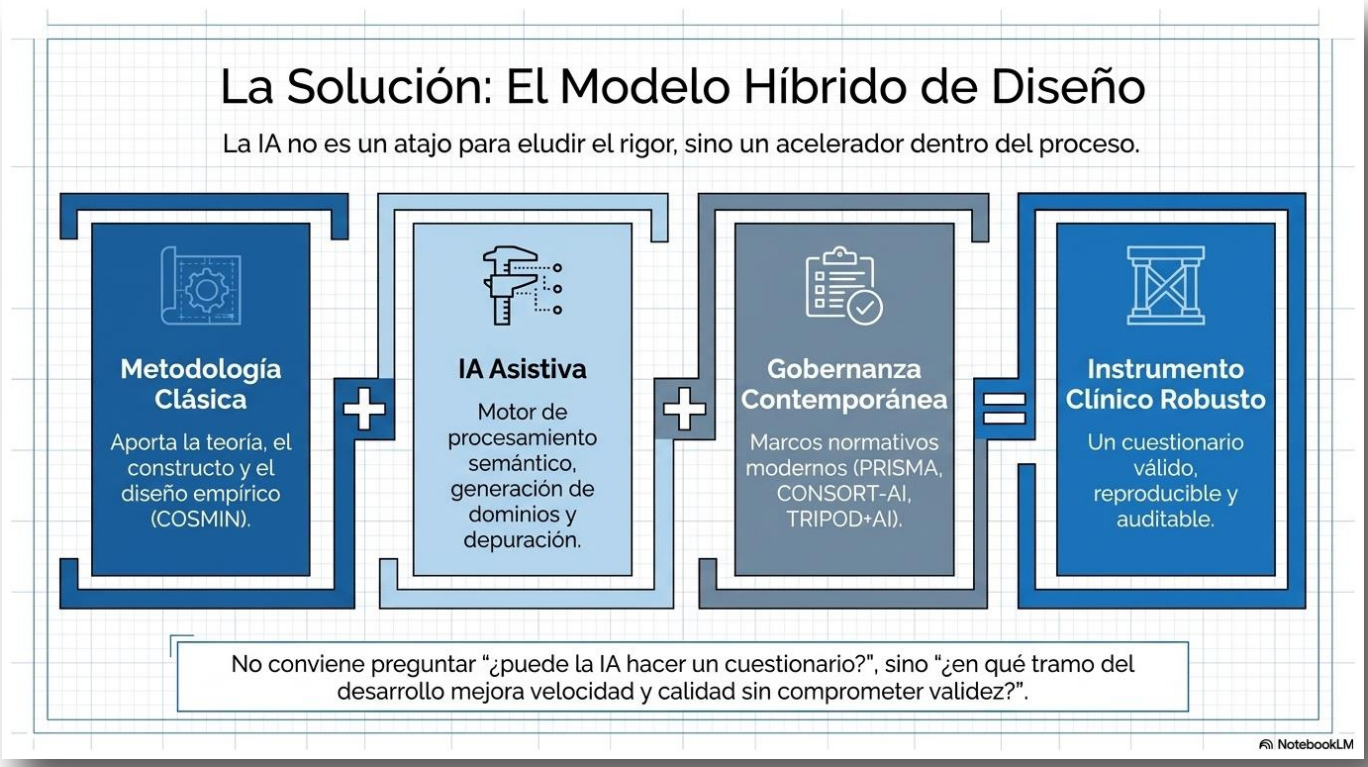
Creencia de que un modelo de lenguaje puede "fabricar" un instrumento "listo para uso" en clínica o investigación.

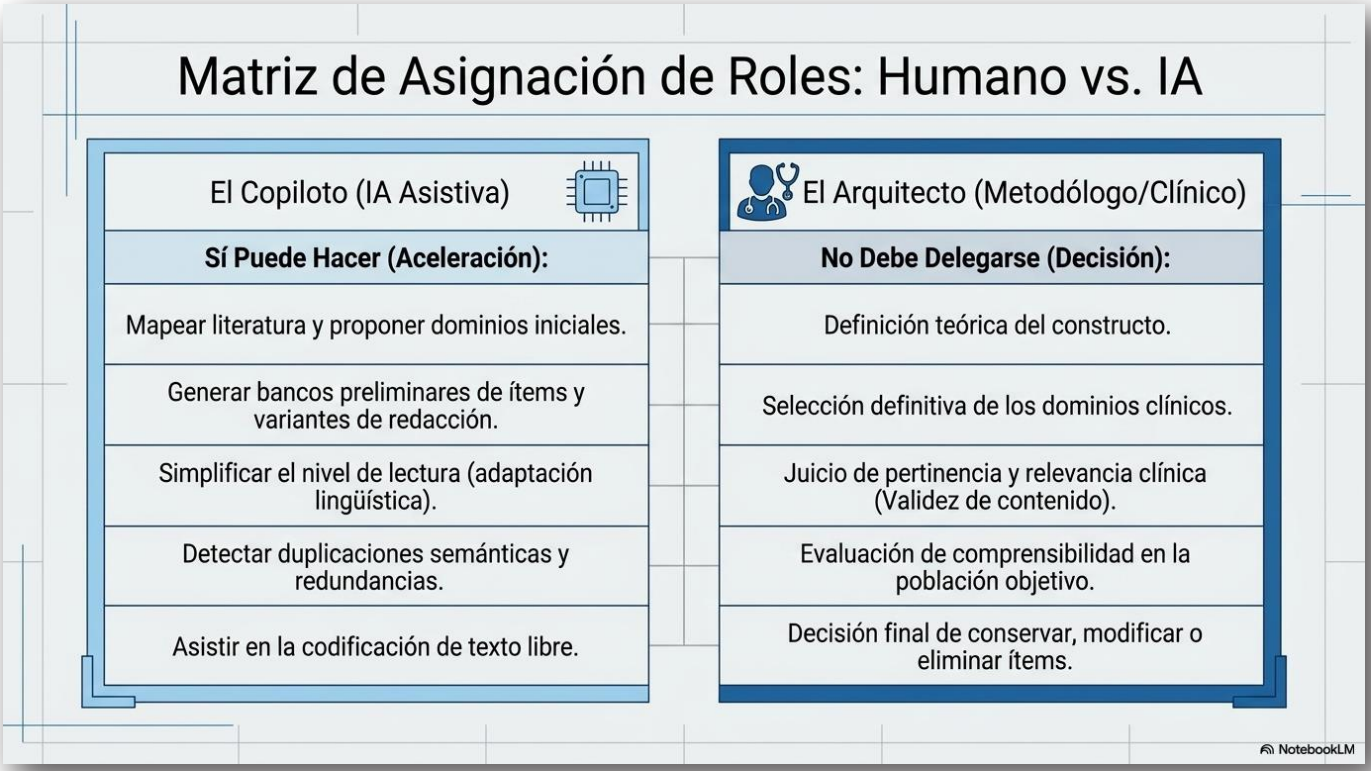
La Realidad Clínica

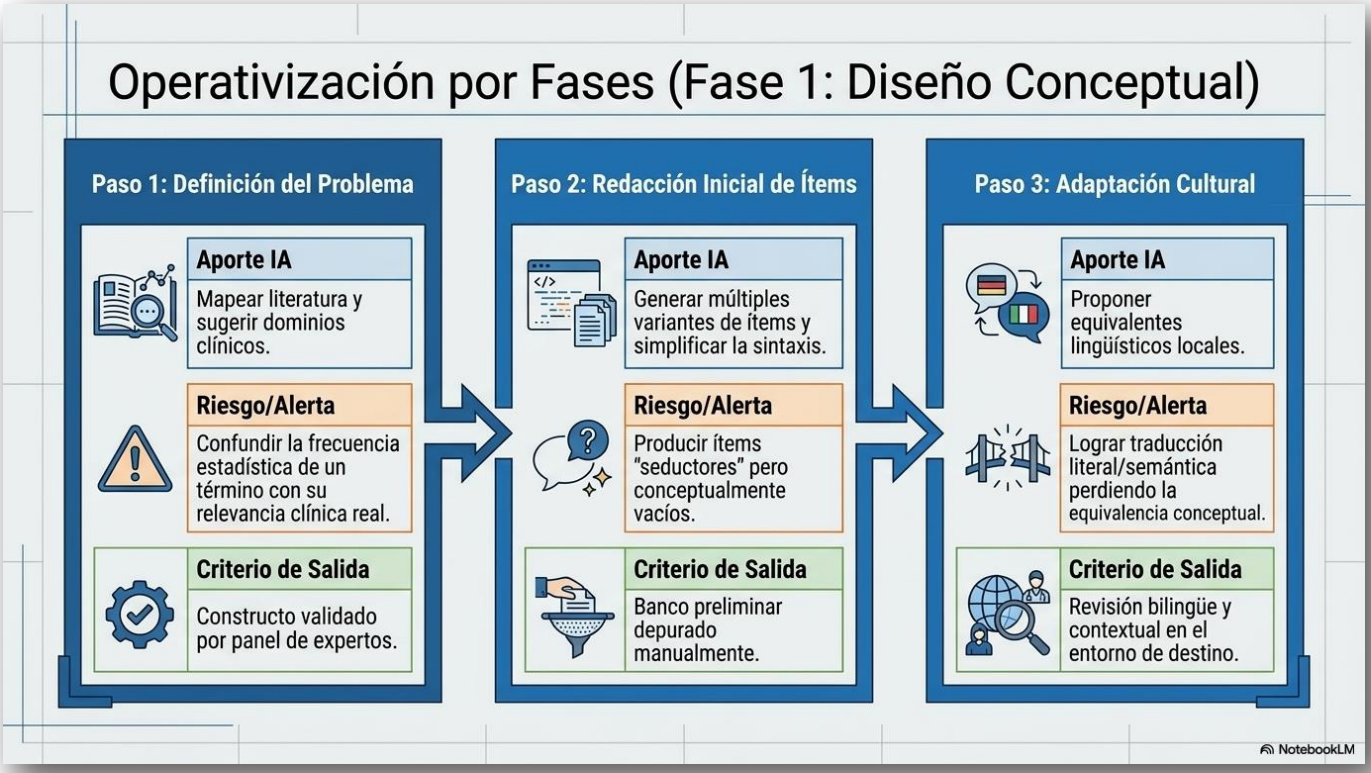
Un cuestionario médico no es una lista de preguntas; es un instrumento de medición.

Su utilidad depende de la definición del constructo, la población objetivo, el contexto y el nivel de error aceptable.

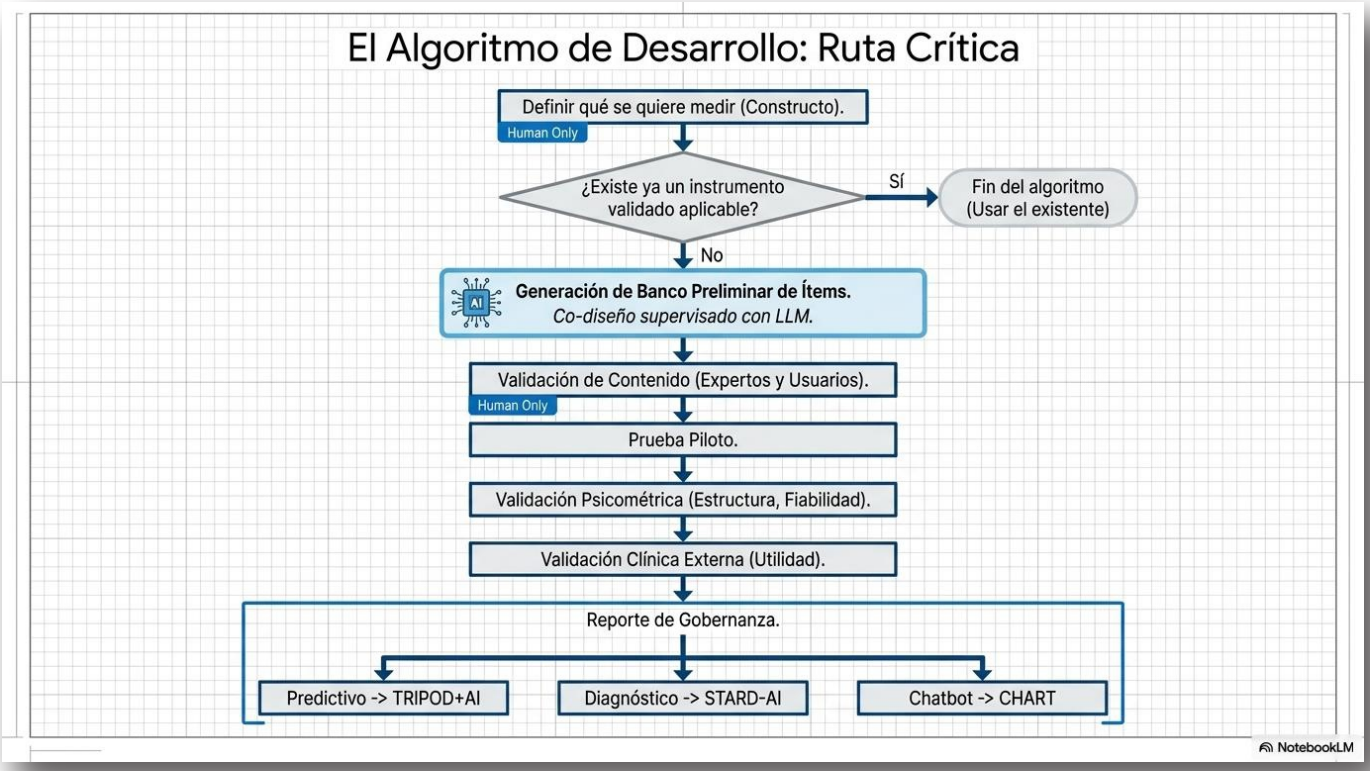

El Principio Rector: La Validez de Contenido Manda
Relevancia
¿Los ítems miden el constructo objetivo en la población correcta?
Validez de Contenido
Exhaustividad
¿Se cubren todos los dominios clínicos necesarios?
Comprensibilidad
¿Lo entienden los pacientes o usuarios reales tal como se pretendía?
El Axioma: Estas tres propiedades requieren evaluación mediante métodos cualitativos y consenso experto humano. La generación automática de texto no puede sustituir el juicio clínico.
NotebookLM


La Solución: El Modelo Híbrido de Diseño
La IA no es un atajo para eludir el rigor, sino un acelerador dentro del proceso.
Metodología Clásica
Aporta la teoría, el constructo y el diseño empírico (COSMIN).
IA Asistiva
Motor de procesamiento semántico, generación de dominios y depuración.
Gobernanza Contemporánea
Marcos normativos modernos (PRISMA, CONSORT-AI, TRIPOD+AI).
Instrumento Clínico Robusto
Un cuestionario válido, reproducible y auditable.
No conviene preguntar "¿puede la IA hacer un cuestionario?", sino "¿en qué tramo del desarrollo mejora velocidad y calidad sin comprometer validez?".
NotebookLM


Matriz de Asignación de Roles: Humano vs. IA
El Copiloto (IA Asistiva)
Sí Puede Hacer (Aceleración):
Mapear literatura y proponer dominios iniciales.
Generar bancos preliminares de ítems y variantes de redacción.
Simplificar el nivel de lectura (adaptación lingüística).
Detectar duplicaciones semánticas y redundancias.
Asistir en la codificación de texto libre.
El Arquitecto (Metodólogo/Clínico)
No Debe Delegarse (Decisión):
Definición teórica del constructo.
Selección definitiva de los dominios clínicos.
Juicio de pertinencia y relevancia clínica (Validez de contenido).
Evaluación de comprensibilidad en la población objetivo.
Decisión final de conservar, modificar o eliminar ítems.
NotebookLM


Operativización por Fases (Fase 1: Diseño Conceptual)
Paso 1: Definición del Problema
Aporte IA
Mapear literatura y sugerir dominios clínicos.
Riesgo/Alerta
Confundir la frecuencia estadística de un término con su relevancia clínica real.
Criterio de Salida
Constructo validado por panel de expertos.
Paso 2: Redacción Inicial de Ítems
Aporte IA
Generar múltiples variantes de ítems y simplificar la sintaxis.
Riesgo/Alerta
Producir ítems "seductores" pero conceptualmente vacíos.
Criterio de Salida
Banco preliminar depurado manualmente.
Paso 3: Adaptación Cultural
Aporte IA
Proponer equivalentes lingüísticos locales.
Riesgo/Alerta
Lograr traducción literal/semántica perdiendo la equivalencia conceptual.
Criterio de Salida
Revisión bilingüe y contextual en el entorno de destino.
NotebookLM


Operativización por Fases (Fase 2: Análisis y Validación)
Paso 4: Lógica del Cuestionario
Aporte IA
Sugerir saltos lógicos y versiones abreviadas.
Riesgo/Alerta
Omitir o mutilar dominios esenciales.
Salida
Flujo de navegación validado clínicamente.
Paso 5: Piloto (Prueba Cognitiva)
Aporte IA
Identificar ambigüedad léxica pre-piloto.
Riesgo/Alerta
Sobreajuste a una muestra piloto pequeña.
Salida
Versión estable tras evaluación en humanos.
Paso 6: Análisis de Datos
Aporte IA
Clasificar respuestas de texto libre, resumir hallazgos cualitativos.
Riesgo/Alerta
Alucinación de patrones, sesgo oculto, pérdida de trazabilidad.
Salida
Auditoría humana estadística completa.
Paso 7: Reporte Final
Aporte IA
Estandarizar descripciones metodológicas.
Riesgo/Alerta
Omitir prompts utilizados, versión del modelo o limitaciones.
Salida
Manuscrito alineado a guías de reporte.
NotebookLM


El Algoritmo de Desarrollo: Ruta Crítica
Definir qué se quiere medir (Constructo).
Human Only
¿Existe ya un instrumento validado aplicable?
Sí
Fin del algoritmo (Usar el existente)
No
AI
Generación de Banco Preliminar de Ítems.
Co-diseño supervisado con LLM.
Validación de Contenido (Expertos y Usuarios).
Human Only
Prueba Piloto.
Validación Psicométrica (Estructura, Fiabilidad).
Validación Clínica Externa (Utilidad).
Reporte de Gobernanza.
Predictivo -> TRIPOD+AI
Diagnóstico -> STARD-AI
Chatbot -> CHART
NotebookLM

Matriz Mínima de Validación (Parte 1: Construcción)

Propiedades psicométricas innegociables según COSMIN 2.0.

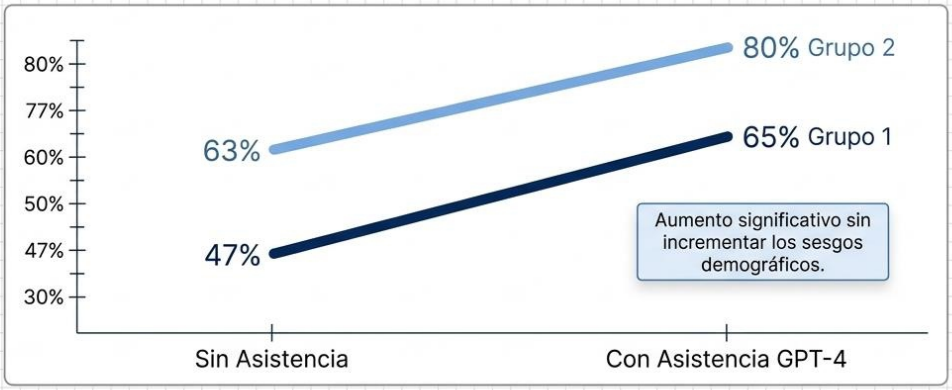
NotebookLM

Matriz Mínima de Validación (Parte 2: Desempeño Clínico)

NotebookLM

Evidencia Clínica: El Impacto de la Colaboración Supervisada

"Aunque la evidencia directa sobre 'diseño de cuestionarios con IA' sigue siendo escasa, existen datos sólidos sobre colaboración humano-IA en tareas clínicas estructuradas."



La IA mejora la capacidad de estructuración y razonamiento solo cuando existe un humano supervisando y aplicando criterios clínicos referenciales (Goh et al., 2025).

NotebookLM

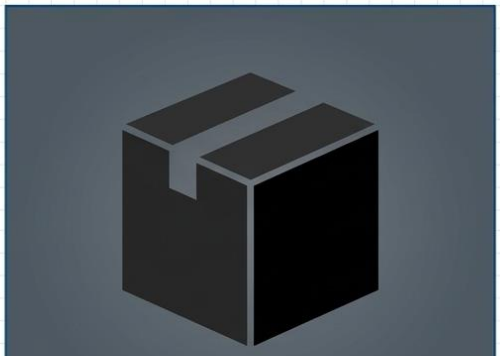
El Ecosistema de Gobernanza: Marcos de Reporte

¿Qué directriz debo utilizar para reportar mi investigación?

Revisiones de Instrumentos (PROMs)	**PRISMA-COSMIN 2024** Marco base: PRISMA 2020
Ensayos Clínicos con componente IA	**CONSORT-AI** Añade 14 ítems específicos
Modelos Predictivos / Clasificación	**TRIPOD+AI**
Precisión Diagnóstica	**STARD-AI**
Cuestionarios Conversacionales (Chatbots)	**CHART** Para consejo en salud

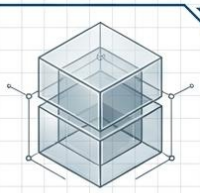
NotebookLM

El Estándar Ético: Evitando la 'Caja Negra' Editorial



El riesgo inminente en la literatura actual es que la metodología del instrumento quede reducido a una 'caja negra' editorial donde es imposible auditar la influencia del modelo de IA. [12,13]

La Solución (RAISE & GAMER)



Transparencia (GAMER)

Obligación estricta de declarar qué herramienta generativa se usó, con qué prompts específicos y en qué versiones de software.

Responsabilidad (RAISE)

Postura conjunta (Cochrane, Campbell, JBI):
- Supervisión humana ininterrumpida.
- Justificación metodológica del uso de IA.
- La IA nunca debe degradar el rigor de la síntesis de evidencia.

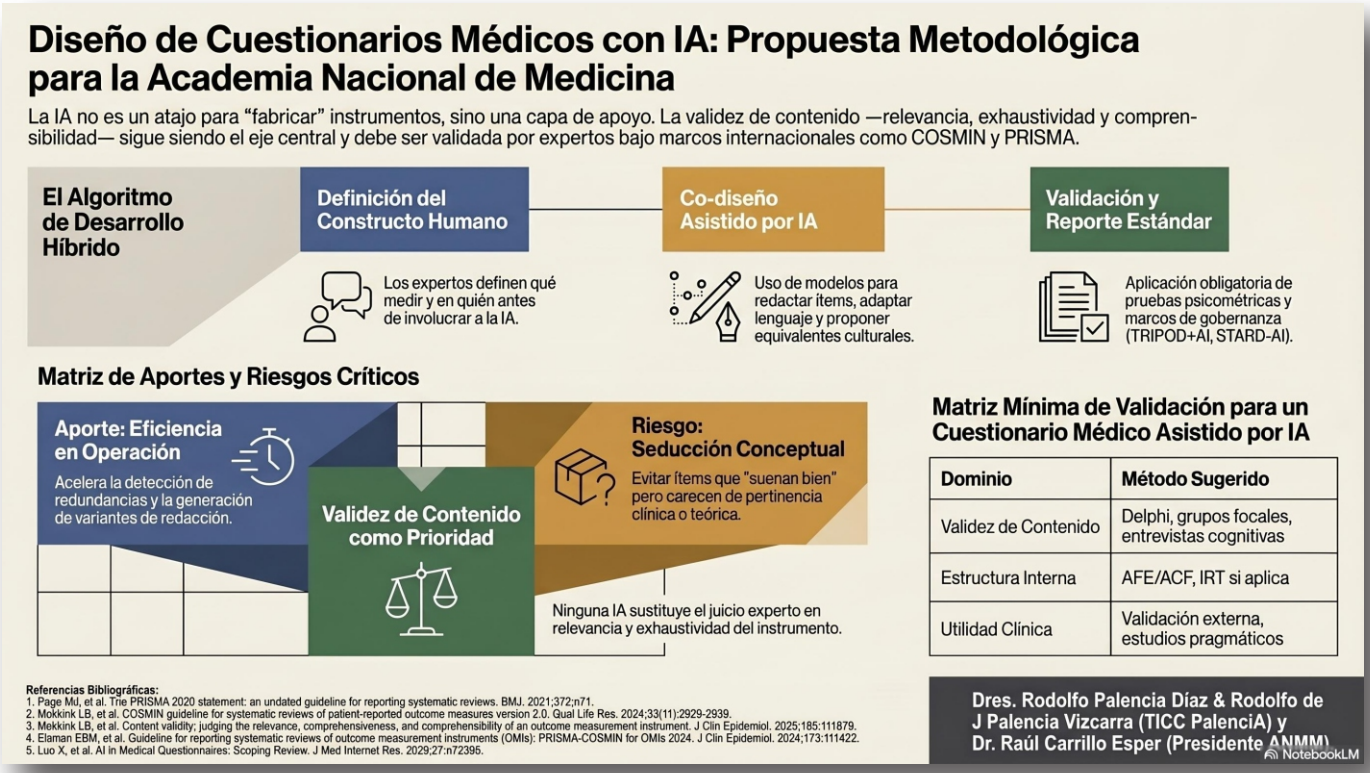
NotebookLM

Conclusión de la Academia

"Un instrumento mal conceptualizado sigue siendo malo aunque lo redacte un gran modelo de lenguaje (LLM). Un instrumento bien conceptualizado puede desarrollarse mejor y más rápido si la IA se usa con disciplina, trazabilidad y validación."

- Definir el constructo clínicamente.
- Co-diseñar ítems bajo supervisión médica.
- Validar psicométrica y clínicamente.
- Reportar con los marcos de gobernanza AI.

NotebookLM



Referencias Bibliográficas (Formato Vancouver)

Page MJ, et al. The PRISMA 2020 statement... BMJ. 2021;372:n71.

Mokkink LB, et al. COSMIN guideline for systematic reviews... Qual Life Res. 2024;33(11):2929-2939.

Mokkink LB, et al. Content validity... a COSMIN perspective. J Clin Epidemiol. 2025;185:111879.

Elsman EBM, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol. 2024;173:111422.

Luo X, et al. AI in Medical Questionnaires: Scoping Review. J Med Internet Res. 2025;27:e72398.

Goh E, et al. Physician clinical decision modification... Commun Med (Lond). 2025;5:59.

Liu X, et al. The CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-e548.

Collins GS, et al. TRIPOD+AI statement... BMJ. 2024;385:e078378.

Sounderajah V, et al. The STARD-AI reporting guideline... Nat Med. 2025;31:3283-3289.

CHART Collaborative. The CHART statement. Artif Intell Med. 2025;168:103222.

Luo X, et al. The GAMER Statement. BMJ Evid Based Med. 2025.

Flemyng E, et al. Position statement on AI use in evidence synthesis... Campbell Syst Rev. 2025;21(4):e70074.

Thomas J, et al. Responsible use of AI in evidence SynthEsis (RAISE) 2... Open Science Framework. 2025.

NotebookLM