

Inteligencia Artificial (IA) en salud:

La era de la caja negra

por Victorino Sepúlveda Arróniz

Nota: el siguiente caso es hipotético e ilustrativo, construido a partir de situaciones documentadas en la literatura sobre implementación de sistemas de IA clínica sin supervisión suficiente. No corresponde a un evento real específico.

El hospital lo llamó Turing Clinical Ops: una IA entrenada con datos de varios centros que pintaba camas en rojo, amarillo o verde en un tablero digital. No explicaba nada; solo mostraba color.

Don Ernesto llegó sin familia y con expediente a medias. Neumonía, oxígeno bajo. El sistema lo marcó verde.

A las 2:17 a.m., empeoró. La enfermera avisó. El residente miró el tablero: verde. Tenía dos camas en rojo parpadeando y una sugerencia discreta: "Optimice recursos: priorice casos de alta probabilidad de beneficio." Corrió hacia las rojas.

A las 3:40 a.m., Don Ernesto entró en paro cardiorrespiratorio.

Al día siguiente, la auditoría mostró el criterio de priorización para el modelo, un expediente incompleto significaba 'bajo riesgo'. No estaba leyendo medicina, sino el formato: campos vacíos, abreviaturas diferentes y datos inconsistentes entre centros. Este es el problema central de la caja negra aplicada al triage clínico: un modelo que influye en decisiones de vida o muerte sin ofrecer ninguna explicación de su razonamiento, imposibilitando la supervisión humana en tiempo real.

Este texto no discute si la IA debe entrar a la medicina (ya entró), sino qué condiciones mínimas deben exigirse para que su adopción no degrade confianza, justicia y seguridad.

Introducción

La medicina contemporánea se encuentra en un punto de inflexión: la complejidad biológica, clínica y social parece superar nuestra capacidad de síntesis. Hoy, un “caso” no es solo un cuerpo: es imagenología, laboratorio seriado, notas clínicas, genética, antecedentes, determinantes sociales, consumo de fármacos, adherencia y un rastro creciente de métricas de bienestar. El problema no es la falta de información; es el exceso, su heterogeneidad y lo complejo de su combinación.

Durante décadas, el avance médico se apoyó en reglas relativamente legibles: signos, escalas, guías, biomarcadores y criterios clínicos interpretables. Pero esa lógica “de manual” tiene límites cuando las señales son débiles, múltiples y contextuales. Ahí es donde la inteligencia artificial (IA) aparece no sólo como oportunidad, sino como una respuesta estadística a una realidad: muchos patrones clínicos existen, pero no son fáciles de verbalizar ni de reducir a una sola variable.

La implementación masiva de IA, sin embargo, trae un debate ético y político inevitable: ¿aceptaremos decisiones y priorizaciones clínicas basadas en modelos que no pueden explicar su razonamiento? El problema no es abstracto. Ya no hablamos solo de algoritmos académicos, sino de productos que llegan a consultorios, laboratorios, aseguradoras y plataformas de atención.

Del hospital al bolsillo

Durante mucho tiempo, el ideal de la medicina fue que una recomendación pudiera justificarse con una cadena de razones: “si A y B, entonces C”. La IA moderna —especialmente el aprendizaje profundo— invierte esa tradición: detecta patrones útiles sin traducirlos a conceptos humanos claros. A estos sistemas se les denomina comúnmente “caja negra”: herramientas que producen resultados precisos pero cuyo proceso interno no puede ser inspeccionado, explicado ni auditado con facilidad, incluso por quienes las diseñaron. Un acontecimiento que ilustra este giro ocurrió en agosto de 2025, cuando la FDA otorgó autorización De Novo a ArteraAI Prostate, software que usa patología digital (y datos clínicos asociados) para pronosticar resultados a largo plazo (Clarke & Saylor, 2025). Más allá de la oncología, el mensaje es amplio: la regulación comienza a aceptar herramientas cuyo valor se demuestra por desempeño y validación, aunque su “lenguaje interno” no sea interpretable. La promesa es real: mejores predicciones, menos variabilidad y apoyo en decisiones complejas. El riesgo también: la opacidad puede esconder sesgos, fallas de generalización o dependencias que no son medicina, sino artefactos del dato.

La discusión, además, ya no se limita a hospitales. En 2026 se acelera otra frontera: la IA como interfaz cotidiana con la salud. OpenAI lanzó ChatGPT Health (Figura 1), que permite conectar datos

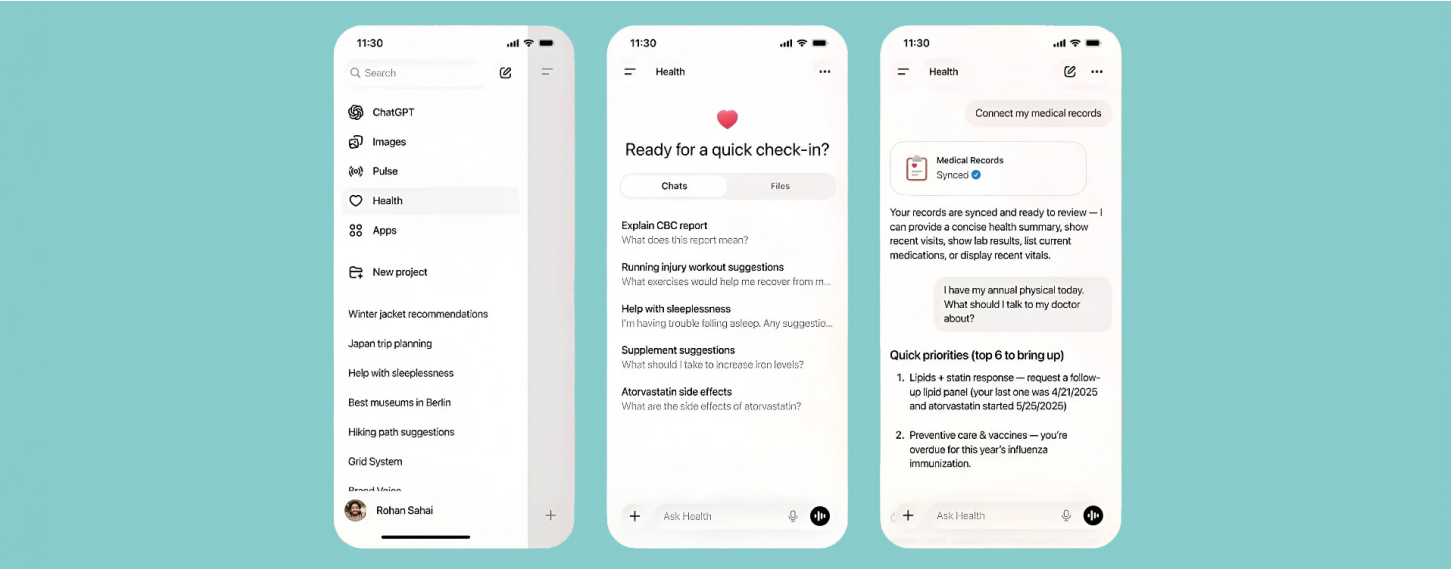


Figura 1. Interfaz de ChatGPT Health. OpenAI (2026).



Imagen Magnifica

como registros médicos, con énfasis declarado en privacidad y control del usuario (OpenAI, 2026). Si bien estas afirmaciones requieren validación independiente y escrutinio regulatorio, la herramienta ilustra una tendencia que ya está normalizando a la IA como “primera ventanilla” para entender síntomas, laboratorios o preparar consultas. Esto reabre preguntas difíciles: ¿Qué significa seguridad con datos tan sensibles?, ¿Qué pasa cuando se confunde orientación con diagnóstico?, ¿Cómo se audita el impacto de un consejo mal entendido?

En paralelo, surgen startups que convierten la mezcla de algoritmos y clínicos en servicio. Lotus Health AI se presenta como atención primaria apoyada por IA y médicos, con “cuidado personalizado” basado en datos y automatización; afirma cumplir

con la HIPAA (Health Insurance Portability and Accountability Act), una normativa estadounidense que exige confidencialidad y medidas de seguridad para el manejo de datos de salud identificables. Además, reporta validación por médicos vinculados a Stanford y Harvard, de forma gratuita (Temkin, 2026). Estas afirmaciones, aunque prometedoras, aún no cuentan con publicaciones revisadas por pares que las respalden de forma independiente, lo que subraya la necesidad de exigir evidencia rigurosa antes de adoptar estos modelos en entornos clínicos formales. Si la IA ya vive en el bolsillo y entra también a la consulta, ¿vamos hacia una salud más accesible y oportuna, o hacia un ecosistema donde la calidad dependa de quién puede pagar, entender términos y conservar el control de su información?

Oportunidades reales

La caja negra no solo exige cautela: también abre oportunidades reales cuando se usa para integrar fuentes de datos que el humano no puede sintetizar. El cáncer es un ejemplo claro: su complejidad biológica va mucho más allá de lo que cualquier clínico puede sintetizar a simple vista. Históricamente, la patología avanzó desde las observaciones microscópicas de Rudolf Virchow en el siglo XIX hasta la identificación de biomarcadores genéticos y proteínas específicas, como el HER2, que revolucionaron el tratamiento en las últimas décadas. Sin embargo, gran parte de los biomarcadores clínicos de mayor impacto fueron identificados entre mediados de los años 60 y principios de los 80; lo que no significa ausencia de progreso, sino que encontrar nuevos marcadores con impacto clínico equivalente se ha vuelto progresivamente más difícil con los métodos convencionales. La realidad es que el cáncer no se define por reglas simples, sino por una “gramática” compleja de señales biológicas débiles que, combinadas, dictan el destino del paciente.

Frente al peligro de la opacidad, la mayor oportunidad reside en la fusión multimodal. El cáncer de mama, por su pronunciada heterogeneidad tumoral, es un ejemplo perfecto de por qué no podemos confiar en una sola fuente de datos. Los modelos que integran imágenes radiológicas, registros clínicos, histopatología y datos genómicos ofrecen predicciones mucho más ricas y fiables que los enfoques unimodales.

La investigación reciente indica que la combinación de datos (fusión híbrida) supera consistentemente a los modelos que solo analizan imágenes o solo genética. Por ejemplo, integrar notas clínicas mediante procesamiento de lenguaje natural con imágenes de ultrasonido ha permitido alcanzar precisiones superiores al 96% en conjuntos de datos específicos y condiciones de laboratorio controladas, aunque estos resultados aún deben ser replicados en poblaciones diversas y entornos clínicos reales antes de generalizarse. Esta capacidad de ver el “cuadro completo” del paciente es lo que podría permitir a la IA operar en la complejidad de la enfermedad, donde los tratamientos estándar suelen fallar (Singh & Patnaik, 2026).

Desafíos de implementación

Aquí está el núcleo del debate en Perspectivas en salud: la IA no falla solo por el modelo, sino por el sistema donde se entrena, se despliega y se usa. Primero, la caja negra complica la rendición de cuentas: si no sabemos por qué recomendó A en lugar de B, la responsabilidad se diluye entre proveedor, hospital, clínico y regulador, erosionando la confianza. Segundo, están los sesgos: un modelo puede rendir peor fuera de la población donde fue entrenado y, peor aún, aprender atajos que reproducen desigualdad (menos datos, menos seguimiento o cierto hospital pueden traducirse en “menor riesgo” sin fundamento clínico). Tercero, la seguridad: conectar expedientes, apps y notas aumenta utilidad, pero también la superficie de riesgo; privacidad no es un checklist, es una decisión social sobre quién custodiará información que puede afectar empleo, seguros o estigma. Y cuarto, la colaboración multicentro: mejora generalización, pero también amplifica errores y mala codificación; no basta “más datos”, hay que demostrar que el modelo aprende medicina, no burocracia.

La IA explicable (XAI) ayuda, pero no basta: muchas explicaciones son aproximaciones. Por eso, la respuesta debe ser diseño por gobernanza: auditorías antes y después del despliegue, trazabilidad de cambios (qué se actualiza, cuándo y con qué evidencia), monitoreo por subgrupos, seguridad y minimización de datos (recolectar lo necesario y controlar accesos) y, sobre todo, responsabilidad clínica explícita cuando una recomendación influye en decisiones como el triage (Hassan et al., 2026).

La pregunta incómoda

¿Estamos dispuestos a aceptar una medicina basada en predicciones más precisas, aunque no comprendamos mecánicamente el “por qué”? La dirección regulatoria y de mercado sugiere un “sí” cauteloso, siempre que haya validación y beneficios demostrables. En México, donde no es raro que médicos y residentes enfrenten guardias que superan las 24 horas, herramientas de IA bien gobernadas podrían ayudar a descargar tareas repetitivas y reducir errores por fatiga. Implementadas con límites claros y supervisión clínica, serían una vía para proteger el bienestar del personal sanitario sin sacrificar la seguridad del paciente.

La IA no debe reemplazar al clínico ni desplazar la responsabilidad hacia un proveedor invisible. Tampoco debe convertir la salud en un flujo de datos explotable por incentivos comerciales opacos. Si herramientas como ChatGPT Health aceleran la adopción, también aceleran el riesgo de normalizar decisiones automatizadas sin debate público, sin auditorías y sin mecanismos de corrección.

La verdadera innovación no será el modelo con mayor precisión en métricas de desempeño, sino el sistema que logre integrar precisión con gobernanza, explicabilidad útil, seguridad real y equidad. La invitación es directa: ¿qué reglas éticas, científicas y políticas podemos exigir para que la IA aumente la medicina sin debilitar la confianza que la sostiene?

Y aquí va la pregunta incómoda: si en el relato inicial quien dijo ‘verde’ hubiera sido un médico o un administrador, ¿lo aceptaríamos igual? La diferencia no es el error; es la firma. Con los humanos hay culpa, debate, aprendizaje y castigo. Con una caja negra solo queda un color y una decisión sin autor. Sin auditoría, no hay confianza; sin confianza, no hay adopción legítima. La responsabilidad de que esa firma exista –y sea exigible– no recae en el algoritmo: recae en los diseñadores que lo construyen, los reguladores que lo aprueban, los clínicos que lo despliegan y la sociedad que decide qué medicina quiere.

Referencias

Clarke, H., & Saylor, B. P. (2025, August 13). FDA grants de novo authorization to ArteraAI Prostate. Urology Times. <https://www.urologytimes.com/view/fda-grants-de-novo-authorization-to-arteraai-prostate>

Hassan, M. M., Tahsin, A., Alam, M. G. R., Alzamil, D., Garg, S., Uddin, M. Z., Choudhury, N., & Fortino, G. (2026). Explainable multimodal fusion for breast carcinoma diagnosis: A systematic review, open problems, and future directions. Computer Methods and Programs in Biomedicine, 274, 109152. <https://doi.org/10.1016/j.cmpb.2025.109152>

OpenAI. (2026, January 7). Introducing ChatGPT Health. <https://openai.com/index/introducing-chatgpt-health/>

Singh, S. K., & Patnaik, K. S. (2026). MammXAI: An XAI integrated adaptive multi-model deep learning approach for breast cancer detection using multi-modality images. Biomedical Signal Processing and Control, 113, 109173. <https://doi.org/10.1016/j.bspc.2025.109173>

Temkin, M. (2026, February 3). Lotus Health nabs \$35M for AI doctor that sees patients for free. TechCrunch. <https://techcrunch.com/2026/02/03/lotus-health-nabs-35m-for-ai-doctor-that-sees-patients-for-free/>



Victorino Sepúlveda Arróniz
Dirección de Investigación VIAC (UEM)

Victorino Sepúlveda Arróniz trabaja en la Universidad de Monterrey como responsable de vinculación y fondos para investigación, enfocándose en la obtención de financiamiento externo para proyectos de investigación. Actualmente cursa un doctorado en Inteligencia Artificial y Optimización en la UANL.