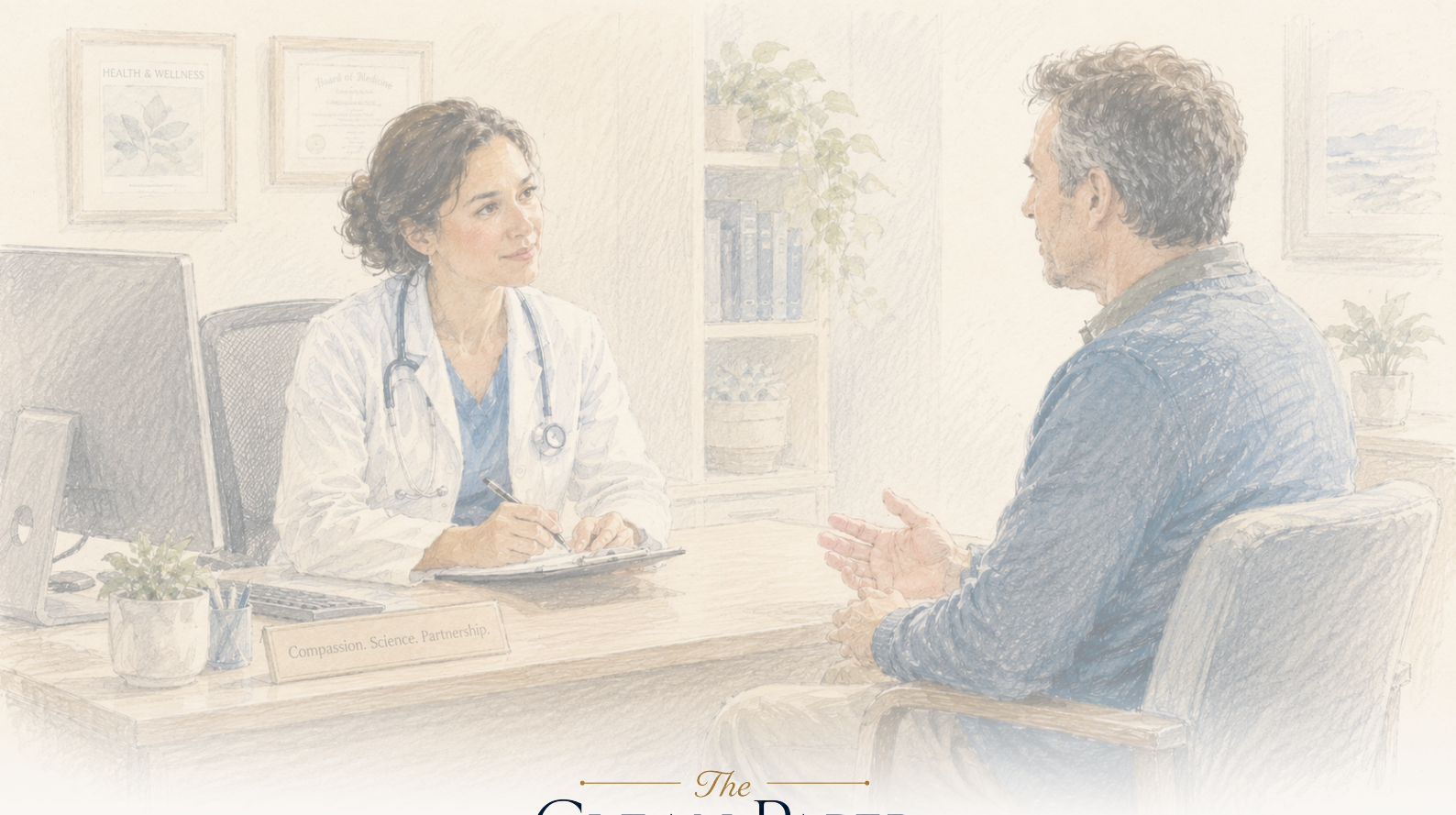


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Una IA clínica que pareció segura y mejoró el papeleo, pero no mejoró los resultados de los pacientes

Un ensayo pragmático, aleatorizado por conglomerados, en 16 centros de atención primaria de Kenia probó una herramienta de apoyo a decisiones con IA generativa añadida al registro usado por clinical officers. Entre 9.691 pacientes, el compuesto estricto a 14 días de fracaso terapéutico adjudicado por expertos fue 2,2 % con la herramienta frente a 2,0 % sin ella (odds ratio ajustado 0,77, IC 95 % 0,55-1,08, P = 0,13): ninguna diferencia significativa. La herramienta no mostró señal de seguridad, mejoró la documentación en todos los dominios evaluados, dejó prescripción y satisfacción sin cambios y tendió a menores costos de antibióticos. No es un estudio fallido; es uno raro y honesto — medido en pacientes, no en benchmarks — y aterriza en la verdad poco glamorosa de que una herramienta que parecía segura y ayudó al proceso no ayudó demostrablemente a los pacientes en dos semanas, y que detectar cualquier beneficio exigiría un ensayo mucho mayor.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial*, Ambrose Agweyu, Paul Mwaniki, Vaishnavi Menon, Robert Korom, Lynda Isaaka, Conrad Wanyama, Xiaoxuan Liu & Bilal A. Mateen (and colleagues), Nature Medicine (2026) · DOI: 10.1038/s41591-026-04503-6

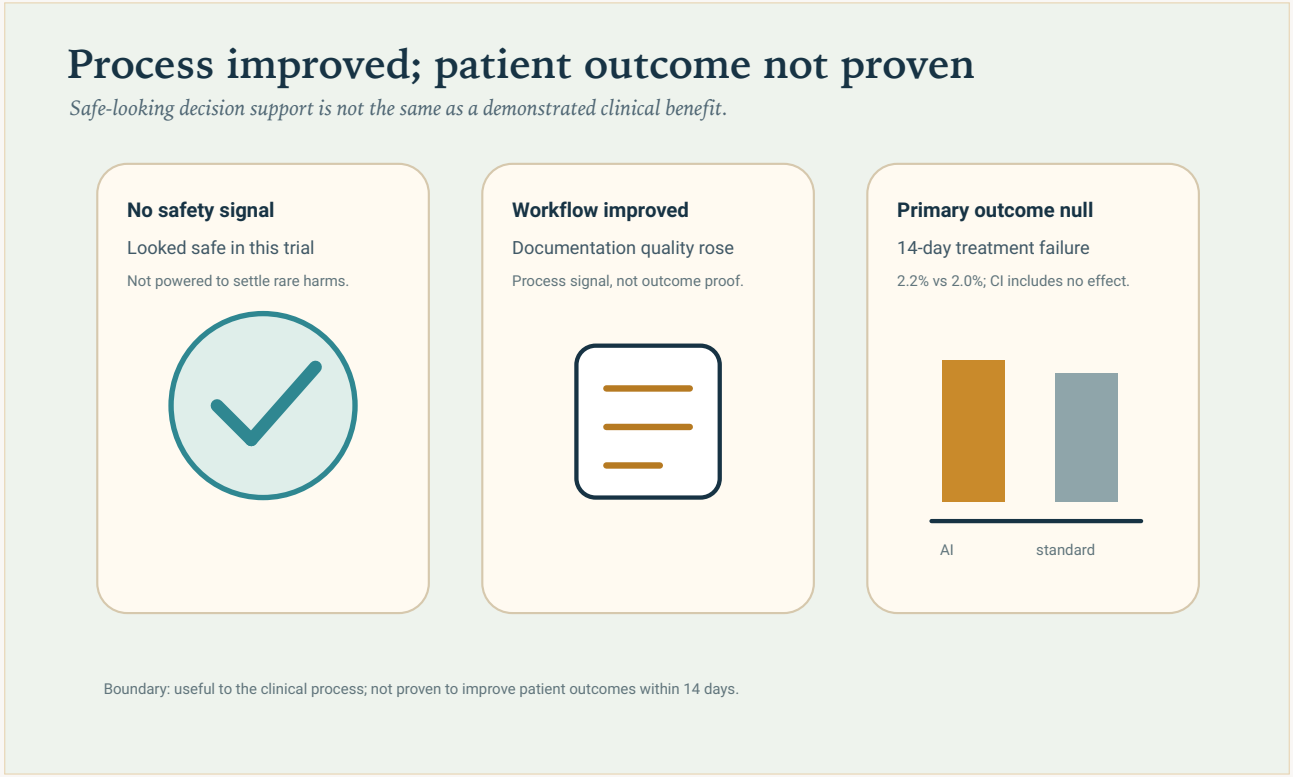
Seguro y útil no es lo mismo que efectivo

La mayoría de los titulares sobre IA médica se construyen a partir del tipo equivocado de estudio. Un modelo puntúa bien en preguntas de examen, o supera a médicos en viñetas seleccionadas, y la historia se escribe sola: la máquina está lista para la clínica.

Este ensayo hizo algo más difícil y raro. Puso una herramienta de apoyo a decisiones con IA generativa en atención primaria real, en centros reales, con pacientes reales, y preguntó lo que de verdad importa: ¿les fue mejor a los pacientes?

La respuesta honesta es no: no de forma medible, no en 14 días. La herramienta no mostró señal de seguridad. Mejoró la calidad de la documentación clínica. Quizá incluso bajó algunos costos de medicamentos. Pero no redujo significativamente los fracasos terapéuticos, y los autores son cuidadosos al decir que cualquier beneficio, si existe, probablemente sea modesto.

Eso no es un fracaso del estudio. Es el estudio funcionando. Así se ve la evidencia responsable sobre IA en medicina cuando se mide en pacientes y no en benchmarks.



La herramienta no mostró señal de seguridad y mejoró la documentación clínica, pero el resultado preespecificado a 14 días en pacientes no mejoró significativamente. Ayuda de proceso no es lo mismo que beneficio probado para el paciente.

Qué hicieron los autores

El equipo realizó un ensayo pragmático, aleatorizado por conglomerados, en **16 centros de atención primaria** operados por una red privada de salud (Penda Health) en los condados de Nairobi y Kiambu, Kenia. La atención en estos centros la prestan en gran medida **clinical officers** — profesionales de nivel medio con un diploma de tres años — a menudo sin acceso fácil a consulta senior.

La unidad de aleatorización fue el clínico, no el paciente. **103 clinical officers** fueron aleatorizados: 52 al brazo de intervención y 51 al brazo control. Ambos brazos usaron el mismo registro médico electrónico en la nube. El brazo de intervención además tuvo **«AI Consult»** (versión 2.0), una herramienta de apoyo a decisiones construida sobre el gran modelo de lenguaje **GPT-4o de OpenAI** e integrada en ese registro. Leía la información documentada por un clínico y podía señalar posibles problemas con el diagnóstico o el plan de tratamiento. Los clínicos mantenían plena autonomía: podían aceptar, modificar o ignorar sus sugerencias.

> **Qué modelo era, y por qué importan los detalles**
>
> La herramienta era **AI Consult 2.0**, ejecutando **GPT-4o de OpenAI** (la versión de mayo de 2025), accedida mediante la API comercial de OpenAI bajo una licencia empresarial y usada con ajustes de baja aleatoriedad (temperatura 0,1). Estaba dentro de un registro electrónico a medida (el EMR de EasyClinic) y guiada por prompts de sistema escritos para alinearse con las guías nacionales de tratamiento de Kenia; los autores publicaron el prompt completo de instrucciones.
> ¿Por qué detallar esto? Porque el resultado trata de *un sistema específico*: una versión de modelo, un prompt, un registro, un entorno; no de «los LLM en medicina» en general. Los autores hacen el mismo punto: llaman a su hallazgo un *benchmark temporal, no una estimación fija de capacidad*. Un modelo más nuevo, un prompt distinto o una clínica menos digitalizada podrían mover el resultado.
> Sobre independencia: OpenAI proporcionó después apoyo en especie (créditos de cómputo en la nube y guía técnica sobre el uso de su API), pero los autores afirman que la decisión de usar OpenAI se tomó *antes* de esa oferta y que OpenAI no tuvo papel en el diseño del ensayo, la recolección de datos, el análisis ni la decisión de publicar.

Entre el 22 de abril y el 16 de julio de 2025, se incluyeron **9.691 pacientes**. El **resultado primario fue deliberadamente centrado en el paciente y estricto**: un *compuesto de fracaso terapéutico* dentro de los 14 días de la visita, adjudicado por expertos; un panel de clínicos, ciego al brazo de estudio, juzgó si cada paciente tuvo un mal resultado como enfermedad no resuelta o empeorada. El ensayo se registró por adelantado (Pan-African Clinical Trials Registry 202502499779176).

Esa elección de diseño es el punto. Es fácil mostrar que una herramienta de IA cambia lo que un clínico escribe. Es mucho más difícil, y mucho más significativo, mostrar que cambia lo que le ocurre al paciente.

Qué encontraron

El resultado primario no mejoró. El fracaso terapéutico ocurrió en 102 de 4.693 pacientes (2,2 %) en el brazo IA y 94 de 4.654 (2,0 %) en el brazo control. Los porcentajes crudos fueron fraccionalmente más altos con IA, pero tras el ajuste la estimación puntual se inclinó hacia beneficio: el **odds ratio ajustado fue 0,77 (intervalo de confianza del 95 % 0,55 a 1,08, P = 0,13)**, no estadísticamente significativo. Ese cambio entre números crudos y ajustados no es un desliz aritmético; el ajuste toma en cuenta diferencias entre los conglomerados de clínicos. En cualquier caso, el intervalo de confianza incluye cómodamente «sin efecto», así que no se puede reclamar beneficio, y en términos absolutos el efecto fue diminuto.

Para una forma sencilla de leer odds ratios, intervalos de confianza y valores P juntos, consulta la guía para leer un resultado clínico.

Ninguna señal de seguridad, dentro de límites. Ningún evento adverso grave fue juzgado relacionado con la herramienta, y una revisión independiente no encontró señal de seguridad. Los autores son honestos sobre el techo de esa tranquilidad: el ensayo no tenía potencia para detectar daños severos raros y no tenía un marco formal preespecificado de no inferioridad o seguridad, así que no puede *probar* seguridad para eventos poco frecuentes.

La documentación mejoró. Entre 2.000 encuentros revisados por expertos ciegos, los clínicos que usaron AI Consult produjeron mejor documentación clínica en todos los dominios evaluados: diagnóstico registrado, plan de tratamiento y completitud general.

La prescripción apenas se movió. No hubo diferencia significativa en prescripción, incluido el uso correcto de antibióticos (odds ratio ajustado 0,86, IC 95 % 0,48 a 1,55). La herramienta no cambió las tasas de prescripción de antibióticos.

Los pacientes no notaron diferencia. Entre 826 pacientes que completaron una encuesta de satisfacción, la satisfacción fue esencialmente idéntica entre brazos, y los tiempos de consulta fueron similares.

Los costos apuntaron ligeramente hacia abajo. En un análisis ajustado, los costos relacionados con antibióticos fueron más bajos en el brazo IA, plausiblemente por elecciones más baratas más que por menos recetas, y el ahorro por paciente en antibióticos pareció superar el costo por paciente de operar la herramienta. Los autores lo señalan como sugerente, no resuelto: una contabilidad completa del costo total de propiedad estaba fuera del ensayo.

El resumen de una línea de los autores es la versión más limpia: la asistencia con LLM fue segura dentro de esos límites, pero no redujo el fracaso terapéutico en 14 días, y cualquier beneficio probablemente sea modesto.

Por qué «sin diferencia significativa» no es «no funciona»

Un resultado primario nulo es fácil de sobreleer en ambas direcciones. Dos cosas detienen la historia simple.

Primero, **el ensayo fue construido para detectar un efecto mayor que el que encontró**. Los malos resultados graves en atención primaria son raros — alrededor de 2 % aquí — así que distinguir un pequeño beneficio real del ruido necesita números enormes. Los cálculos de potencia post-hoc de los autores sugieren que detectar un efecto del tamaño observado requeriría **un ensayo mucho más grande, del orden de más de 100.000 pacientes**. Un resultado no significativo en un estudio de este tamaño no descarta un pequeño beneficio real; significa que este estudio no pudo resolverlo.

Segundo, **la comparación estuvo parcialmente borrosa**. Un error de configuración dio brevemente a algunos clínicos del brazo control acceso a AI Consult, y los clínicos de una red compartida hablan entre sí y llevan hábitos a través de la frontera. Ambos efectos tienden a hacer que los dos brazos se parezcan más, empujando cualquier diferencia real hacia cero. Además, la red anfitriona ya funcionaba con estándares relativamente altos, lo que deja menos espacio para que una herramienta muestre mejora.

Nada de esto rescata un titular de «avance». Pero sí significa que la lectura correcta es calibrada, no deflacionaria: en el endpoint más duro y honesto, esta herramienta no ayudó demostrablemente a los pacientes en dos semanas, aunque sí ayudó mediblemente al registro y pareció segura.

Qué no prueba esto

- **No** muestra que la IA mejorara los resultados de los pacientes. En el endpoint primario a 14 días, no hubo beneficio significativo.
- **No** muestra que la IA sea inútil. La estimación puntual la favoreció, la documentación mejoró en todos los dominios y los costos de medicamentos tendieron a bajar; el resultado nulo es compatible con un pequeño beneficio real que el ensayo fue demasiado pequeño para confirmar.
- **No** prueba que la herramienta sea segura para daños raros. No mostró señal de seguridad, pero no tenía potencia ni diseño para certificar seguridad frente a eventos severos poco comunes.
- **No** muestra que «la IA vence a los médicos» ni reemplaza a clínicos. Es apoyo a decisiones; el clínico conservó plena autoridad para aceptarlo o rechazarlo.
- **No** se generaliza automáticamente. El ensayo se realizó en una sola red privada urbana de Kenia; entornos rurales, periurbanos y de ingresos más altos podrían diferir en ambas direcciones.
- **No** establece ahorro de costos. La señal de costos es sugerente, no una evaluación económica completa.

Qué tan fuerte es la evidencia

Para la afirmación central — ninguna reducción probada del daño al paciente a corto plazo — la evidencia es fuerte *como diseño* y adecuadamente humilde *como conclusión*. Un ensayo prospectivo, prerregistrado, aleatorizado por conglomerados, con un resultado compuesto a nivel de paciente y ciego, está cerca de la mejor evidencia real que se puede reunir para una herramienta como esta. Es mucho más informativo que una puntuación de benchmark o un estudio de viñetas.

Para los hallazgos secundarios — mejor documentación, prescripción sin cambios, satisfacción similar, menores costos de antibióticos — la evidencia es buena pero debe leerse como secundaria: señales de apoyo, no el titular, y vulnerables a las mismas limitaciones de contaminación y red única.

Para seguridad, la evidencia es tranquilizadora pero acotada: no se encontró señal, pero no era un estudio construido para hallar daños raros.

La postura más útil no es «funciona» ni «fracasó». Es: un ensayo cuidadoso en el mundo real encontró que esta herramienta de IA no levantó señal de seguridad y mejoró el proceso de atención, sin demostrar beneficio en resultados de pacientes en dos semanas, y que detectar cualquier beneficio así requeriría un estudio mucho mayor.

Por qué importa

El debate sobre IA médica está hambriento exactamente de este tipo de evidencia. Hay miles de artículos que muestran modelos aprobando exámenes y emparejando a clínicos en casos ordenados. Hay muy pocos ensayos grandes, pragmáticos y aleatorizados que midan si a pacientes reales les va mejor. Este es uno de ellos, y aterriza en la verdad poco glamorosa: pasar el examen no es lo mismo que ayudar al paciente.

Esa brecha es toda la historia. Una herramienta puede ser genuinamente útil para clínicos — notas más claras, menores facturas de medicamentos, un segundo par de ojos — y aun así no mover un resultado duro de paciente en quince días. Ambas cosas pueden ser ciertas a la vez, y un sistema de salud maduro tiene que sostenerlas juntas en lugar de elegir la conveniente.

También reajusta la carga de prueba en una dirección útil. Si una empresa quiere afirmar que su IA clínica mejora la atención, la evidencia relevante no es una tabla de clasificación. Es un ensayo como este, sobre resultados que importan a los pacientes, e idealmente uno mayor, porque la lección honesta aquí es que los beneficios modestos necesitan estudios grandes para verse.

Resumen limpio

Un ensayo pragmático, aleatorizado por conglomerados, en 16 centros de atención primaria de Kenia probó una herramienta de apoyo a decisiones con IA generativa («AI Consult») añadida al registro

electrónico usado por clinical officers. Entre 9.691 pacientes, el compuesto de fracaso terapéutico adjudicado por expertos dentro de 14 días fue 2,2 % con la herramienta frente a 2,0 % sin ella (odds ratio ajustado 0,77, IC 95 % 0,55-1,08, P = 0,13): ninguna diferencia significativa. La herramienta no mostró señal de seguridad, mejoró la documentación clínica en todos los dominios evaluados, no cambió la prescripción, dejó la satisfacción del paciente sin cambios y se asoció con costos de antibióticos algo menores. Los autores concluyen que fue segura dentro de esos límites, pero no redujo el fracaso terapéutico, con cualquier beneficio probablemente modesto; detectar un efecto del tamaño observado requeriría un ensayo mucho mayor (del orden de 100.000 pacientes). El resultado no muestra que la IA clínica mejore los resultados de los pacientes, ni que sea inútil: muestra que una herramienta que parecía segura y ayudó al proceso no ayudó demostrablemente a los pacientes en dos semanas, en una red privada urbana.

No-BS check

Lo que muestra el artículo: En un ensayo aleatorizado del mundo real, añadir una herramienta de apoyo a decisiones basada en LLM a registros de atención primaria no levantó señal de seguridad, mejoró la calidad de la documentación clínica, no cambió la prescripción y no redujo significativamente un compuesto estricto de fracaso terapéutico a 14 días.

Lo que es plausible pero no está probado: Que la herramienta produzca una pequeña reducción real del fracaso terapéutico demasiado pequeña para que este ensayo la detecte; que ahorre dinero una vez contados todos los costos; que mejor documentación eventualmente se traduzca en mejor atención.

Lo que no muestra: Que la IA clínica mejore los resultados de los pacientes; que sea segura frente a eventos raros; que reemplace o supere a clínicos; que estos resultados se transfieran a entornos rurales o de ingresos más altos; que el ahorro de costos esté establecido.

Principales limitaciones: Potencia para un efecto mayor que el observado (los resultados raros necesitan muestras muy grandes); un error de configuración contaminó el brazo control y los hábitos de una red compartida borronean la comparación, ambos sesgando hacia no diferencia; una sola red privada urbana con estándares ya altos; sin marco preespecificado de no inferioridad o seguridad; horizonte corto de 14 días.

¿Cuánta confianza debería tener un lector general? Alta en que la herramienta no mostró señal de seguridad en este ensayo y mejoró la documentación. Alta en que no mejoró demostrablemente aquí los resultados de pacientes a 14 días. Baja para cualquier afirmación de que «funciona» o «fracasa» como intervención de beneficio para pacientes: esa pregunta sigue genuinamente abierta y necesita un estudio mucho mayor. Postura adecuada: un resultado cuidadoso del mundo real sobre una herramienta que no levantó señal de seguridad y mejoró el proceso de atención, no un veredicto de que la IA transforma — o destruye — la atención primaria.

Sources

Agweyu et al., Nature Medicine (2026)
https://doi.org/10.1038/s41591-026-04503-6

Pan-African Clinical Trials Registry, registration 202502499779176
https://pactr.samrc.ac.za/

EurekAlert / PATH press release on the trial (2026)
https://www.eurekalert.org/news-releases/1133583