The Clean Paper

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

Un'AI clinica apparsa sicura e capace di migliorare la documentazione — ma senza migliorare gli esiti dei pazienti

Un trial pragmatico, cluster-randomizzato, in 16 strutture di cure primarie keniote ha testato uno strumento di supporto decisionale basato su AI generativa aggiunto alla cartella usata dai clinical officers. Tra 9.691 pazienti, il rigoroso composito a 14 giorni di fallimento terapeutico giudicato da esperti era 2,2% con lo strumento contro 2,0% senza (odds ratio aggiustato 0,77, IC 95% 0,55–1,08, P = 0,13) — nessuna differenza significativa. Lo strumento non ha mostrato segnali di sicurezza, ha migliorato la documentazione in tutti i domini valutati, ha lasciato prescrizione e soddisfazione invariate e tendeva a costi antibiotici più bassi. Non è uno studio fallito; è uno studio raro e onesto — misurato sui pazienti, non sui benchmark — e arriva alla verità poco glamour che uno strumento apparso sicuro e utile al processo non ha aiutato dimostrabilmente i pazienti in due settimane, e che rilevare un eventuale beneficio richiederebbe un trial molto più grande.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · 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

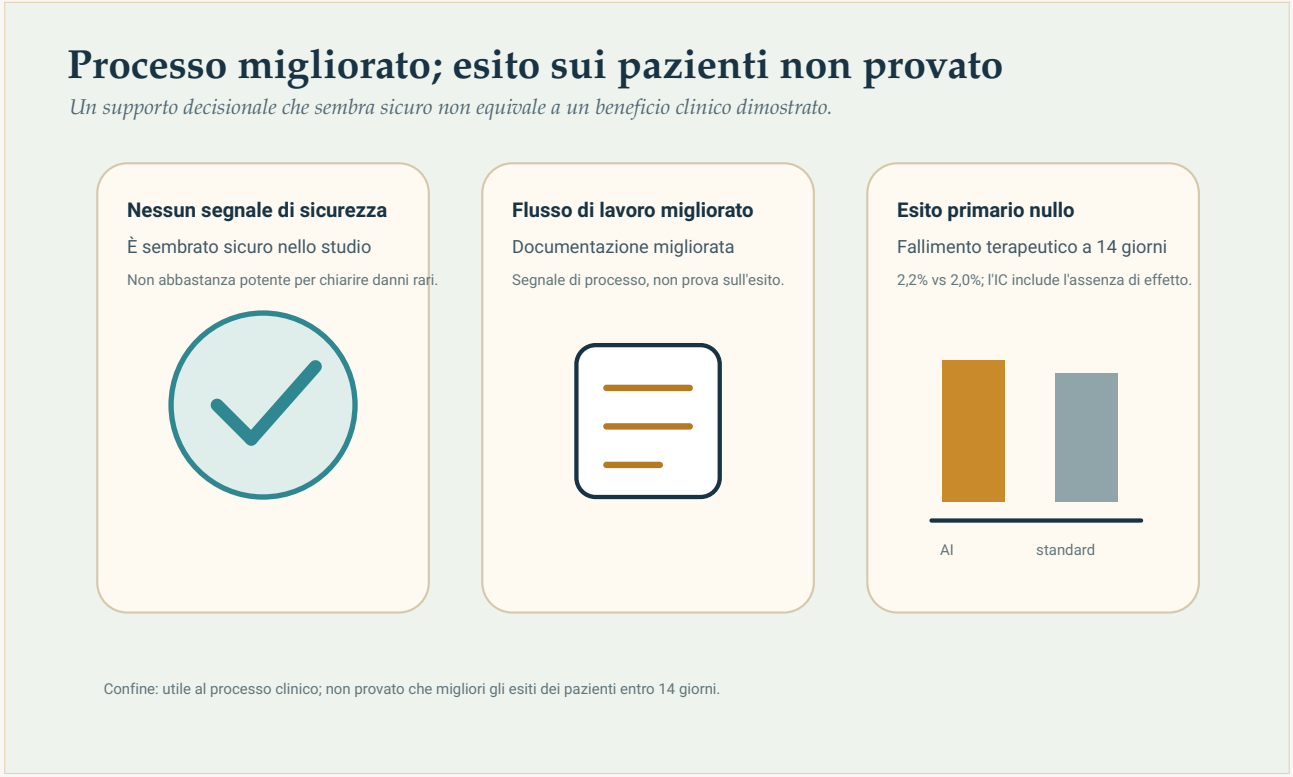
Sicuro e utile non è la stessa cosa di efficace

La maggior parte dei titoli sull'AI medica nasce dal tipo sbagliato di studio. Un modello va bene su domande d'esame, o batte i medici su vignette curate, e la storia si scrive da sola: la macchina è pronta per la clinica.

Questo trial ha fatto qualcosa di più difficile e più raro. Ha messo uno strumento di supporto decisionale basato su AI generativa nella medicina di base reale, in strutture reali, con pazienti reali, e ha posto la domanda che conta davvero: i pazienti sono stati meglio?

La risposta onesta è no — non in modo misurabile, non in 14 giorni. Lo strumento non ha mostrato segnali di sicurezza. Ha migliorato la qualità della documentazione clinica. Potrebbe perfino aver abbassato alcuni costi dei farmaci. Ma non ha ridotto in modo significativo i fallimenti terapeutici, e gli autori dicono con cautela che qualunque beneficio, se esiste, è probabilmente modesto.

Questo non è un fallimento dello studio. È lo studio che funziona. È così che si presentano prove responsabili sull'AI in medicina quando viene misurata sui pazienti invece che sui benchmark.



Lo strumento non ha mostrato segnali di sicurezza e ha migliorato la documentazione clinica, ma l'esito prespecificato sui pazienti a 14 giorni non è migliorato in modo significativo. Aiutare il processo non è la stessa cosa di dimostrare un beneficio per il paziente.

Che cosa hanno fatto gli autori

Il team ha condotto un trial pragmatico, cluster-randomizzato, in **16 strutture di cure primarie** gestite da una rete sanitaria privata (Penda Health) nelle contee di Nairobi e Kiambu, in Kenya. In queste strutture l'assistenza è fornita in gran parte da **clinical officers** — professionisti di livello intermedio con un diploma triennale — spesso senza facile accesso a una consulenza senior.

L'unità di randomizzazione era il clinico, non il paziente. **103 clinical officers** sono stati randomizzati: 52 nel braccio intervento e 51 nel braccio controllo. Entrambi i bracci usavano la stessa cartella clinica elettronica cloud. Il braccio intervento aveva in più **"AI Consult"** (versione 2.0), uno strumento di supporto decisionale costruito sul large language model **GPT-4o di OpenAI** e incorporato in quella cartella. Leggeva le informazioni documentate dal clinico e poteva segnalare possibili problemi nella diagnosi o nel piano terapeutico. I clinici mantenevano piena autonomia: potevano accettare, modificare o ignorare i suggerimenti.

> **Quale modello era, e perché i dettagli contano**
>
> Lo strumento era **AI Consult 2.0**, basato su **GPT-4o di OpenAI** (release di maggio 2025), raggiunto tramite l'API commerciale di OpenAI sotto licenza enterprise e usato con impostazioni a bassa casualità (temperatura 0,1). Era dentro una cartella elettronica su misura (l'EMR EasyClinic) ed era guidato da system prompt scritti per allinearsi alle linee guida terapeutiche nazionali keniote; gli autori hanno pubblicato il prompt completo.
> Perché esplicitarlo? Perché il risultato riguarda *un sistema specifico* — una versione di modello, un prompt, una cartella, un contesto — non "gli LLM in medicina" in generale. Gli autori fanno lo stesso punto: chiamano il risultato un *benchmark temporale più che una stima fissa di capacità*. Un modello più nuovo, un prompt diverso o una clinica meno digitalizzata potrebbero spostare l'esito.
> Sull'indipendenza: OpenAI ha poi fornito supporto in-kind (crediti di cloud compute e guida tecnica sull'uso della sua API), ma gli autori dichiarano che la decisione di usare OpenAI era stata presa *prima* di quell'offerta e che OpenAI non ha avuto ruolo nel disegno del trial, nella raccolta dati, nell'analisi o nella decisione di pubblicare.

Tra il 22 aprile e il 16 luglio 2025 sono stati arruolati **9.691 pazienti**. L'**esito primario era deliberatamente centrato sul paziente e severo**: un *composito di fallimento terapeutico* entro 14 giorni dalla visita, giudicato da esperti — un panel di clinici ha valutato, in cieco rispetto al braccio di studio, se ogni paziente avesse avuto un esito negativo come malattia non risolta o peggiorata. Il trial era registrato in anticipo (Pan-African Clinical Trials Registry 202502499779176).

Questa scelta di disegno è il punto. È facile mostrare che uno strumento AI cambia ciò che un clinico scrive. È molto più difficile, e molto più significativo, mostrare che cambia ciò che succede al paziente.

Che cosa hanno trovato

L'esito primario non è migliorato. Il fallimento terapeutico si è verificato in 102 su 4.693 pazienti (2,2%) nel braccio AI e in 94 su 4.654 (2,0%) nel braccio controllo. Le percentuali grezze erano frazionalmente più alte con l'AI, ma dopo aggiustamento la stima puntuale pendeva verso il beneficio: l'**odds ratio aggiustato era 0,77 (intervallo di confidenza 95% 0,55–1,08, P = 0,13)** — non statisticamente significativo. Questo ribaltamento tra numeri grezzi e aggiustati non è un errore aritmetico; l'aggiustamento tiene conto delle differenze tra i cluster di clinici. In ogni caso l'intervallo di confidenza include comodamente "nessun effetto", quindi non si può rivendicare un beneficio, e in termini assoluti l'effetto era minuscolo.

Per un modo in linguaggio semplice di leggere insieme odds ratio, intervalli di confidenza e P-value, vedi la guida ai risultati clinici.

Nessun segnale di sicurezza — entro limiti. Nessun evento avverso grave è stato giudicato correlato allo strumento, e una revisione indipendente non ha trovato segnali di sicurezza. Gli autori sono onesti sul tetto di questa rassicurazione: il trial non aveva potenza per rilevare danni severi rari e non aveva un piano prespecificato di non inferiorità o di valutazione formale della sicurezza, quindi non può *provare* la sicurezza per eventi non comuni.

La documentazione è migliorata. Tra 2.000 incontri valutati da esperti in cieco, i clinici che usavano AI Consult hanno prodotto documentazione clinica migliore in tutti i domini valutati — diagnosi registrata, piano terapeutico e completezza complessiva.

La prescrizione si è mossa appena. Non c'è stata differenza significativa nella prescrizione, incluso l'uso corretto di antibiotici (odds ratio aggiustato 0,86, IC 95% 0,48–1,55). Lo strumento non ha cambiato i tassi di prescrizione di antibiotici.

I pazienti non hanno notato differenze. Tra 826 pazienti che hanno completato un sondaggio di soddisfazione, la soddisfazione era sostanzialmente identica tra i bracci e i tempi di consultazione erano simili.

I costi puntavano leggermente verso il basso. In un'analisi aggiustata, i costi legati agli antibiotici erano più bassi nel braccio AI — plausibilmente attraverso scelte più economiche più che meno prescrizioni — e il risparmio per paziente sugli antibiotici sembrava superare il costo per paziente di far girare lo strumento. Gli autori lo segnalano come suggestivo, non risolto: una contabilità completa del costo totale di possesso era fuori dal trial.

La sintesi in una riga degli autori è la versione più pulita: l'assistenza LLM era sicura entro quei limiti ma non ha ridotto il fallimento terapeutico entro 14 giorni, e qualunque beneficio è probabilmente modesto.

Perché “nessuna differenza significativa” non è “non funziona”

Un esito primario nullo è facile da sovra-leggere in entrambe le direzioni. Due cose fermano la storia semplice.

Primo, **il trial era costruito per catturare un effetto più grande di quello trovato**. Gli esiti seri negativi in cure primarie sono rari — circa il 2% qui — quindi distinguere un piccolo beneficio reale dal rumore richiede numeri enormi. I calcoli post-hoc di potenza degli autori suggeriscono che rilevare un effetto della dimensione osservata richiederebbe un **trial molto più grande, dell’ordine di oltre 100.000 pazienti.** Un risultato non significativo in uno studio di questa dimensione non esclude un piccolo beneficio reale; significa che questo studio non poteva risolverlo.

Secondo, **il confronto era in parte sfocato**. Un errore di configurazione ha dato brevemente ad alcuni clinici del braccio controllo accesso ad AI Consult, e i clinici in una rete condivisa parlano tra loro e portano abitudini oltre il confine. Entrambi gli effetti tendono a rendere i due bracci più simili, spingendo qualunque differenza reale verso zero. Inoltre, la rete ospitante lavorava già a standard relativamente alti, lasciando meno spazio a uno strumento per mostrare miglioramenti.

Tutto questo non salva un titolo da “svolta”. Ma significa che la lettura corretta è calibrata, non deflazionaria: sull’endpoint più duro e onesto, questo strumento non ha aiutato dimostrabilmente i pazienti in due settimane — mentre ha aiutato in modo misurabile la tenuta della cartella ed è apparso sicuro.

Che cosa questo studio *non* dimostra

- **Non** mostra che l’AI abbia migliorato gli esiti dei pazienti. Sull’endpoint primario a 14 giorni non c’è stato beneficio significativo.
- **Non** mostra che l’AI sia inutile. La stima puntuale la favoriva, la documentazione è migliorata in tutti i domini e i costi dei farmaci tendevano al ribasso; il risultato nullo è compatibile con un piccolo beneficio reale che il trial era troppo piccolo per confermare.
- **Non** prova che lo strumento sia sicuro per danni rari. Non ha mostrato segnali di sicurezza, ma non era dimensionato o disegnato per certificare la sicurezza per eventi severi non comuni.
- **Non** mostra che “l’AI batte i medici” o sostituisca i clinici. È supporto decisionale; il clinico conservava piena autorità di accettarlo o respingerlo.
- **Non** generalizza automaticamente. Il trial si è svolto in una singola rete privata urbana in Kenya; contesti rurali, periurbani o ad alto reddito potrebbero differire in entrambe le direzioni.
- **Non** stabilisce risparmi di costo. Il segnale sui costi è suggestivo, non una valutazione economica completa.

Quanto sono solide le prove?

Per l'affermazione centrale — nessuna riduzione dimostrata del danno a breve termine per i pazienti — le prove sono solide *per il disegno dello studio* e appropriatamente caute *nelle conclusioni*. Un trial prospettico, preregistrato, cluster-randomizzato, con esito composito a livello paziente giudicato in cieco, è vicino alle migliori prove reali che si possano raccogliere per uno strumento di questo tipo. È molto più informativo di un punteggio su benchmark o di uno studio su vignette.

Per i risultati secondari — documentazione migliore, prescrizione invariata, soddisfazione simile, costi antibiotici più bassi — le prove sono buone ma vanno lette come secondarie: segnali di supporto, non il titolo, e vulnerabili agli stessi limiti di contaminazione e singola rete.

Per la sicurezza, le prove sono rassicuranti ma circoscritte: nessun segnale trovato, ma non uno studio costruito per trovare danni rari.

La posizione più utile non è né "funziona" né "ha fallito". È: un trial reale e cauto ha trovato che questo strumento AI non ha sollevato segnali di sicurezza e ha migliorato il processo di cura, senza dimostrare un beneficio sugli esiti dei pazienti in due settimane — e rilevare un eventuale beneficio richiederebbe uno studio molto più grande.

Perché conta

Il dibattito sull'AI medica ha fame esattamente di questo tipo di evidenza. Esistono migliaia di lavoro che mostrano modelli superare esami e uguagliare clinici su casi ordinati. Esistono pochissimi trial grandi, pragmatici e randomizzati che misurano se pazienti reali stanno meglio. Questo è uno di quelli, e arriva alla verità poco glamour: passare il test non è la stessa cosa che aiutare il paziente.

Quel divario è tutta la storia. Uno strumento può essere davvero utile ai clinici — note più chiare, bollette dei farmaci più basse, un secondo paio d'occhi — e comunque non spostare un endpoint duro del paziente in quindici giorni. Entrambi i fatti possono essere veri insieme, e un sistema sanitario maturo deve tenerli insieme invece di scegliere quello comodo.

Reimposta anche l'onere della prova in una direzione utile. Se un'azienda vuole affermare che la sua AI clinica migliora la cura, la prova rilevante non è una classifica. È un trial come questo, su esiti che contano per i pazienti — e, idealmente, uno più grande, perché la lezione onesta qui è che benefici modesti richiedono studi grandi per essere visti.

In breve

Un trial pragmatico, cluster-randomizzato, in 16 strutture di cure primarie keniote ha testato uno strumento di supporto decisionale basato su AI generativa ("AI Consult") aggiunto alla cartella elettronica usata dai clinical officers. Tra 9.691 pazienti, il composito di fallimento terapeutico entro 14

giorni, giudicato da esperti, era 2,2% con lo strumento contro 2,0% senza (odds ratio aggiustato 0,77, IC 95% 0,55–1,08, P = 0,13) — nessuna differenza significativa. Lo strumento non ha mostrato segnali di sicurezza, ha migliorato la documentazione clinica in tutti i domini valutati, non ha cambiato la prescrizione, ha lasciato invariata la soddisfazione dei pazienti ed era associato a costi antibiotici un po' più bassi. Gli autori concludono che era sicuro entro quei limiti ma non ha ridotto il fallimento terapeutico, con qualunque beneficio probabilmente modesto; rilevare un effetto della dimensione osservata richiederebbe un trial molto più grande, dell'ordine di 100.000 pazienti. Il risultato non mostra che l'AI clinica migliori gli esiti dei pazienti, né che sia inutile — mostra che uno strumento apparso sicuro e utile al processo non ha aiutato dimostrabilmente i pazienti in due settimane, in una rete privata urbana.

Valutazione critica

Che cosa mostra il lavoro: In un trial randomizzato nel mondo reale, aggiungere uno strumento di supporto decisionale basato su LLM alle cartelle di cure primarie non ha sollevato segnali di sicurezza, ha migliorato la qualità della documentazione clinica, non ha cambiato la prescrizione e non ha ridotto in modo significativo un composito rigoroso di fallimento terapeutico dei pazienti a 14 giorni.

Che cosa è plausibile ma non dimostrato: Che lo strumento produca una piccola riduzione reale del fallimento terapeutico troppo piccola perché questo trial la rilevi; che faccia risparmiare denaro una volta contati tutti i costi; che una documentazione migliore si traduca alla fine in una cura migliore.

Che cosa non mostra: Che l'AI clinica migliori gli esiti dei pazienti; che sia insicura per eventi rari; che sostituisca o superi i clinici; che questi risultati si trasferiscano a contesti rurali o ad alto reddito; che il risparmio di costo sia stabilito.

Limiti principali: Potenza tarata su un effetto più grande di quello osservato (gli esiti rari richiedono campioni molto grandi); un errore di configurazione ha contaminato il braccio controllo e le abitudini di una rete condivisa sfocano il confronto, entrambi spingendo verso nessuna differenza; una singola rete privata urbana con standard già alti; nessun piano prespecificato di non inferiorità o di valutazione della sicurezza; orizzonte breve di 14 giorni.

Quanta fiducia dovrebbe avere un lettore generale? Alta che lo strumento non abbia mostrato segnali di sicurezza in questo trial e abbia migliorato la documentazione. Alta che qui non abbia migliorato dimostrabilmente gli esiti dei pazienti a 14 giorni. Bassa per qualunque affermazione che "funziona" o "fallisce" come intervento di beneficio per i pazienti — quella domanda è davvero irrisolta e richiede uno studio molto più grande. Posizione appropriata: un risultato reale e cauto su uno strumento che non ha sollevato segnali di sicurezza e ha migliorato il processo di cura, non un verdetto che l'AI trasformi — o distrugga — la medicina di base.

Fonti

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