The Clean Paper

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

Uma IA clínica que pareceu segura e melhorou a documentação — mas não melhorou os desfechos dos doentes

Um ensaio pragmático, aleatorizado por clusters, em 16 unidades de cuidados de saúde primários no Quénia testou uma ferramenta de apoio à decisão baseada em IA generativa adicionada ao processo clínico eletrónico usado por clinical officers. Entre 9 691 doentes, o rigoroso compósito de falha terapêutica em 14 dias, adjudicado por peritos, foi de 2,2% com a ferramenta contra 2,0% sem ela (odds ratio ajustado 0,77; IC 95% 0,55–1,08; P = 0,13) — sem diferença significativa. A ferramenta não mostrou qualquer sinal de segurança, melhorou a documentação em todos os domínios avaliados, deixou prescrição e satisfação inalteradas e apontou para custos mais baixos com antibióticos. Não é um estudo falhado; é um estudo raro e honesto — medido em doentes, não em benchmarks — e chega à verdade pouco glamorosa de que uma ferramenta que pareceu segura e ajudou o processo não ajudou de forma demonstrável os doentes em duas semanas, e que detetar um eventual benefício exigiria um ensaio muito maior.

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

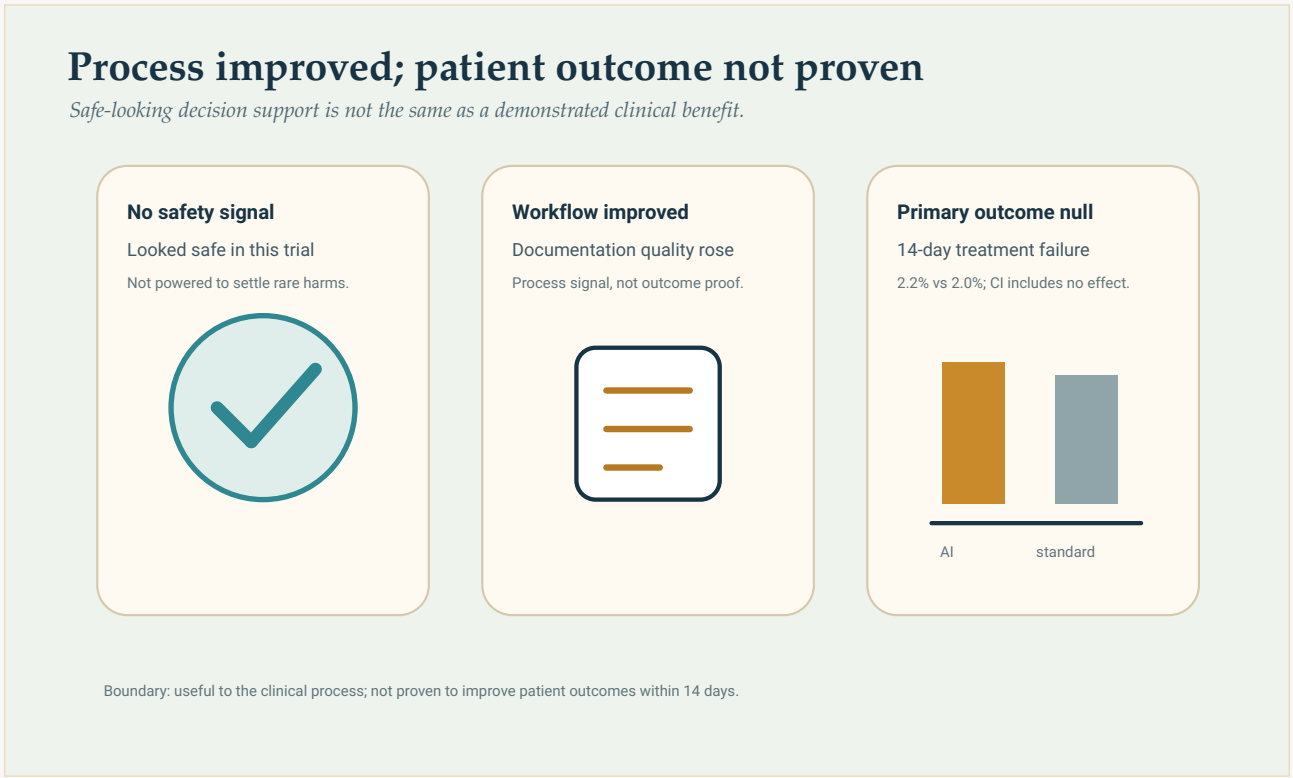
Seguro e útil não é o mesmo que eficaz

A maioria das manchetes sobre IA médica nasce do tipo errado de estudo. Um modelo obtém bons resultados em perguntas de exame ou supera médicos em vinhetas clínicas cuidadosamente selecionadas, e a história escreve-se quase sozinha: a máquina está pronta para a clínica.

Este ensaio fez algo mais difícil e muito mais raro. Colocou uma ferramenta de apoio à decisão baseada em IA generativa em cuidados de saúde primários reais, em unidades reais, com doentes reais, e fez a pergunta que realmente interessa: os doentes ficaram melhor?

A resposta honesta é não — não de forma mensurável, em 14 dias. A ferramenta não mostrou qualquer sinal de segurança preocupante. Melhorou a qualidade da documentação clínica. Pode até ter reduzido alguns custos com medicamentos. Mas não reduziu significativamente as falhas terapêuticas, e os autores são cuidadosos ao dizer que qualquer benefício, se existir, será provavelmente modesto.

Isto não é uma falha do estudo. É o estudo a funcionar. É assim que se parece evidência responsável sobre IA em medicina quando se mede o que acontece aos doentes, em vez de pontuações em benchmarks.



A ferramenta não mostrou qualquer sinal de segurança e melhorou a documentação clínica, mas o desfecho pré-especificado dos doentes aos 14 dias não melhorou significativamente. Ajudar o processo não é o mesmo que demonstrar benefício para o doente.

O que fizeram os autores

A equipa realizou um ensaio pragmático, aleatorizado por clusters, em **16 unidades de cuidados de saúde primários** operadas por uma rede privada de saúde (Penda Health) nos condados de Nairobi e Kiambu, no Quénia. Nestas unidades, os cuidados são prestados sobretudo por **clinical officers** — profissionais de nível intermédio com um diploma de três anos — muitas vezes sem acesso fácil a consulta com profissionais mais experientes.

A unidade de aleatorização foi o clínico, não o doente. Foram aleatorizados **103 clinical officers**: 52 para o braço de intervenção e 51 para o braço de controlo. Ambos os braços utilizavam o mesmo processo clínico eletrónico baseado na cloud. O braço de intervenção dispunha ainda do **«AI Consult»** (versão 2.0), uma ferramenta de apoio à decisão construída sobre o grande modelo de linguagem **GPT-4o da OpenAI** e integrada nesse processo. A ferramenta lia a informação registada pelo clínico e podia assinalar possíveis problemas no diagnóstico ou no plano terapêutico. Os clínicos mantinham autonomia total: podiam aceitar, modificar ou ignorar as sugestões.

> **Que modelo era, e porque importam os pormenores**
>
> A ferramenta era o **AI Consult 2.0**, executando o **GPT-4o da OpenAI** (versão de maio de 2025), acedido através da API comercial da OpenAI ao abrigo de uma licença empresarial e configurado com baixa aleatoriedade (temperatura 0,1). Estava integrado num processo clínico eletrónico desenvolvido à medida (o EMR EasyClinic) e era orientado por prompts de sistema escritos para o alinhar com as normas nacionais de tratamento do Quénia; os autores publicaram integralmente o prompt de instruções.
> Porque explicitar tudo isto? Porque o resultado diz respeito a *um sistema específico* — uma versão de modelo, um prompt, um processo clínico eletrónico e uma configuração — e não aos «LLM na medicina» em geral. Os próprios autores fazem a mesma ressalva: chamam ao resultado um *benchmark temporal, e não uma estimativa fixa da capacidade*. Um modelo mais recente, um prompt diferente ou uma clínica menos digitalizada poderiam alterar o resultado.
> Quanto à independência: a OpenAI forneceu posteriormente apoio em espécie (créditos de computação na cloud e orientação técnica sobre a utilização da API), mas os autores afirmam que a decisão de usar a OpenAI foi tomada *antes* dessa oferta e que a OpenAI não teve qualquer papel no desenho do ensaio, recolha de dados, análise ou decisão de publicar.

Entre 22 de abril e 16 de julho de 2025 foram incluídos **9 691 doentes**. O **desfecho primário foi deliberadamente centrado no doente e exigente**: um *compósito de falha terapêutica* dentro de 14 dias após a consulta, adjudicado por peritos — um painel de clínicos avaliou, sem saber a que braço pertencia cada doente, se tinha ocorrido um mau desfecho, como doença não resolvida ou agravada. O ensaio foi registado antecipadamente (Pan-African Clinical Trials Registry 202502499779176).

É precisamente essa escolha de desenho que interessa. É fácil mostrar que uma ferramenta de IA altera aquilo que um clínico escreve. É muito mais difícil — e muito mais significativo — mostrar que altera aquilo que acontece ao doente.

O que encontraram

O desfecho primário não melhorou. Ocorreu falha terapêutica em 102 de 4 693 doentes (2,2%) no braço de IA e em 94 de 4 654 (2,0%) no braço de controlo. As percentagens brutas eram ligeiramente superiores com IA, mas, após ajustamento, a estimativa pontual inclinou-se para um possível benefício: o **odds ratio ajustado foi 0,77 (intervalo de confiança de 95% 0,55 a 1,08; P = 0,13)** — não estatisticamente significativo. Esta inversão entre os números brutos e os ajustados não é um erro aritmético; o ajustamento tem em conta diferenças entre os clusters de clínicos. De qualquer modo, o intervalo de confiança inclui confortavelmente «nenhum efeito», pelo que não se pode afirmar benefício e, em termos absolutos, o efeito foi minúsculo.

Para uma explicação em linguagem simples de como ler em conjunto odds ratios, intervalos de confiança e valores de P, veja o guia para interpretar resultados clínicos.

Nenhum sinal de segurança — dentro dos limites do estudo. Nenhum acontecimento adverso grave foi considerado relacionado com a ferramenta e uma revisão independente não encontrou qualquer sinal de segurança. Os autores são transparentes sobre o limite desta tranquilização: o ensaio não tinha poder estatístico para detetar danos graves raros e não possuía um quadro pré-especificado de não inferioridade ou de segurança formal, pelo que não pode *demonstrar* segurança para acontecimentos pouco frequentes.

A documentação melhorou. Entre 2 000 consultas analisadas por peritos cegos quanto ao grupo, os clínicos que utilizaram o AI Consult produziram melhor documentação clínica em todos os domínios avaliados — diagnóstico registado, plano terapêutico e completude global.

A prescrição quase não mudou. Não houve diferença significativa na prescrição, incluindo na utilização correta de antibióticos (odds ratio ajustado 0,86; IC 95% 0,48 a 1,55). A ferramenta não alterou as taxas de prescrição de antibióticos.

Os doentes não notaram diferença. Entre os 826 doentes que responderam a um inquérito de satisfação, a satisfação foi praticamente idêntica nos dois braços e a duração das consultas foi semelhante.

Os custos apontaram ligeiramente para baixo. Numa análise ajustada, os custos relacionados com antibióticos foram menores no braço de IA — plausivelmente por escolhas mais baratas, e não por menos prescrições — e a poupança em antibióticos por doente pareceu superar o custo por doente da utilização da ferramenta. Os autores assinalam este resultado como sugestivo, não estabelecido: uma análise completa do custo total de propriedade estava fora do âmbito do ensaio.

O resumo dos próprios autores numa frase é a versão mais limpa: a assistência por LLM mostrou-se segura dentro destes limites, mas não reduziu a falha terapêutica em 14 dias, e qualquer benefício é provavelmente modesto.

Porque «nenhuma diferença significativa» não significa «não funciona»

É fácil interpretar em excesso um desfecho primário nulo em qualquer das direções. Há duas razões para evitar a história simples.

Primeiro, **o ensaio foi desenhado para detetar um efeito maior do que aquele que encontrou.** Maus desfechos graves em cuidados primários são raros — aqui, cerca de 2% — e distinguir um benefício pequeno e real do ruído exige números enormes. Os cálculos de poder estatístico pós-hoc dos próprios autores sugerem que detetar um efeito do tamanho observado exigiria um **ensaio muito maior, da ordem de mais de 100 000 doentes**. Um resultado não significativo num estudo desta dimensão não exclui um benefício pequeno e real; significa que este estudo não o conseguiu resolver.

Segundo, **a comparação ficou parcialmente esbatida.** Um erro de configuração deu durante algum tempo acesso ao AI Consult a alguns clínicos do braço de controlo, e profissionais de uma mesma rede falam entre si e transportam hábitos para lá da fronteira entre grupos. Ambos os efeitos tendem a tornar os dois braços mais semelhantes, empurrando uma eventual diferença real em direção a zero. Além disso, a rede em que o estudo decorreu já funcionava com padrões relativamente elevados, deixando menos espaço para a ferramenta demonstrar melhoria.

Nada disto salva uma manchete de «avanço revolucionário». Mas significa que a leitura correta deve ser calibrada, não depreciativa: no endpoint mais difícil e mais honesto, esta ferramenta não ajudou de forma demonstrável os doentes em duas semanas — embora tenha ajudado de forma mensurável a documentação e tenha parecido segura.

O que isto *não* demonstra

- **Não** mostra que a IA melhorou os desfechos dos doentes. No endpoint primário aos 14 dias, não houve benefício significativo.
- **Não** mostra que a IA seja inútil. A estimativa pontual favoreceu-a, a documentação melhorou de forma consistente e os custos dos medicamentos tenderam a diminuir; o resultado nulo é compatível com um pequeno benefício real que o ensaio era demasiado pequeno para confirmar.
- **Não** demonstra que a ferramenta seja segura para danos raros. Não mostrou qualquer sinal de segurança, mas o estudo não tinha poder nem desenho para certificar segurança em acontecimentos graves pouco frequentes.
- **Não** mostra que «a IA vence os médicos» ou substitui os clínicos. Trata-se de *apoio* à decisão; o clínico manteve autoridade total para aceitar ou rejeitar as sugestões.
- **Não** se generaliza automaticamente. O ensaio decorreu numa única rede urbana privada no Quénia; contextos rurais, periurbanos ou de países com rendimentos mais elevados podem diferir em qualquer direção.
- **Não** estabelece poupanças de custos. O sinal económico é sugestivo, não uma avaliação económica concluída.

Quão forte é a evidência?

Para a afirmação central — nenhuma redução demonstrada de dano a curto prazo para o doente — a evidência é forte *no desenho* e adequadamente prudente *na conclusão*. Um ensaio prospetivo, pré-registado, aleatorizado por clusters, com um desfecho composto ao nível do doente e adjudicação cega, aproxima-se da melhor evidência do mundo real que se pode obter para uma ferramenta deste tipo. É muito mais informativo do que uma pontuação num benchmark ou um estudo com vinhetas clínicas.

Para os resultados secundários — melhor documentação, prescrição inalterada, satisfação semelhante e menores custos com antibióticos — a evidência é boa, mas deve ser lida como secundária: sinais de apoio, não a manchete, e sujeitos às mesmas limitações de contaminação entre grupos e de uma única rede.

Quanto à segurança, a evidência é tranquilizadora, mas limitada: não foi encontrado qualquer sinal, porém não era um estudo concebido para encontrar danos raros.

A posição mais útil não é «funciona» nem «falhou». É esta: um ensaio cuidadoso no mundo real descobriu que esta ferramenta de IA não levantou sinais de segurança e melhorou o processo de cuidados, sem demonstrar benefício nos desfechos dos doentes em duas semanas — e detetar um eventual benefício exigiria um estudo muito maior.

Porque importa

O debate sobre IA médica carece precisamente deste tipo de evidência. Há milhares de artigos a mostrar modelos a dominar exames e a igualar clínicos em casos cuidadosamente preparados. Existem muito poucos ensaios grandes, pragmáticos e aleatorizados que meçam se os doentes reais ficam melhor. Este é um deles, e chega a uma verdade pouco glamorosa: passar no teste não é o mesmo que ajudar o doente.

Essa distância é toda a história. Uma ferramenta pode ser genuinamente útil para os clínicos — notas mais claras, contas de medicamentos mais baixas, um segundo par de olhos — e ainda assim não alterar um desfecho clínico exigente em duas semanas. Ambos os factos podem ser verdade ao mesmo tempo, e um sistema de saúde maduro tem de os manter juntos em vez de escolher o mais conveniente.

O estudo também redefine, de forma útil, o ónus da prova. Se uma empresa quiser afirmar que a sua IA clínica melhora os cuidados, a evidência relevante não é uma tabela classificativa. É um ensaio como este, com desfechos que importam aos doentes — e, idealmente, um ensaio maior, porque a lição honesta aqui é que benefícios modestos exigem grandes estudos para serem detetados.

Resumo claro

Um ensaio pragmático, aleatorizado por clusters, realizado em 16 unidades de cuidados de saúde primários no Quénia testou uma ferramenta de apoio à decisão baseada em IA generativa («AI Consult») adicionada ao processo clínico eletrónico usado por clinical officers. Entre 9 691 doentes, o compósito de falha terapêutica em 14 dias, adjudicado por peritos, ocorreu em 2,2% com a ferramenta contra 2,0% sem ela (odds ratio ajustado 0,77; IC 95% 0,55–1,08; P = 0,13) — sem diferença significativa. A ferramenta não mostrou qualquer sinal de segurança, melhorou a documentação clínica em todos os domínios avaliados, não alterou a prescrição, deixou inalterada a satisfação dos doentes e associou-se a custos de antibióticos ligeiramente menores. Os autores concluem que foi segura dentro destes limites, mas não reduziu a falha terapêutica, sendo qualquer benefício provavelmente modesto; detetar um efeito do tamanho observado exigiria um ensaio muito maior (da ordem dos 100 000 doentes). O resultado não mostra que a IA clínica melhore os desfechos dos doentes, nem que seja inútil — mostra que uma ferramenta que pareceu segura e ajudou o processo não ajudou de forma demonstrável os doentes em duas semanas, numa única rede urbana privada.

Verificação sem exageros

O que o artigo mostra: Num ensaio aleatorizado no mundo real, adicionar uma ferramenta de apoio à decisão baseada num LLM aos processos clínicos de cuidados primários não levantou qualquer sinal de segurança, melhorou a qualidade da documentação clínica, não alterou a prescrição e não reduziu significativamente um compósito rigoroso de falha terapêutica aos 14 dias.

O que é plausível, mas não está demonstrado: Que a ferramenta produza uma pequena redução real na falha terapêutica, demasiado pequena para este ensaio detetar; que poupe dinheiro quando todos os custos são contabilizados; que melhor documentação venha a traduzir-se em melhores cuidados.

O que não mostra: Que a IA clínica melhore os desfechos dos doentes; que seja segura para acontecimentos raros; que substitua ou supere os clínicos; que estes resultados se transfiram para contextos rurais ou de maior rendimento; que a poupança de custos esteja estabelecida.

Principais limitações: O estudo tinha poder para um efeito maior do que o observado (desfechos raros exigem amostras muito grandes); um erro de configuração contaminou o braço de controlo e hábitos partilhados na rede esbatem a comparação, ambos enviesando em direção à ausência de diferença; uma única rede urbana privada com padrões já elevados; sem quadro pré-especificado de não inferioridade ou segurança; horizonte curto de 14 dias.

Quanta confiança deve ter um leitor geral? Elevada de que a ferramenta não mostrou qualquer sinal de segurança neste ensaio e melhorou a documentação. Elevada de que não melhorou de forma demonstrável os desfechos dos doentes aos 14 dias neste contexto. Baixa para qualquer afirmação de que «funciona» ou «falha» enquanto intervenção capaz de beneficiar o doente — essa pergunta continua genuinamente por resolver e exige um estudo muito maior. A posição adequada é vê-lo como um resultado cuidadoso do mundo real sobre uma ferramenta que não levantou sinais de segurança e melhorou o processo de cuidados, não como um veredicto de que a IA transforma —

ou destrói — os cuidados de saúde primários.

Fontes

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