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 pacientes

Um ensaio pragmático, randomizado por conglomerados, em 16 unidades de atenção primária no Quênia testou uma ferramenta de apoio à decisão com IA generativa adicionada ao prontuário usado por clinical officers. Entre 9.691 pacientes, o composto estrito de falha de tratamento em 14 dias adjudicado por especialistas foi 2,2% com a ferramenta versus 2,0% sem ela (razão de chances ajustada 0,77, IC 95% 0,55-1,08, P = 0,13) — sem diferença significativa. A ferramenta não mostrou sinal de segurança, melhorou a documentação em todos os domínios avaliados, deixou prescrição e satisfação inalteradas e tendeu a custos menores com antibióticos. Isso não é um estudo fracassado; é um estudo raro e honesto — medido em pacientes, não em benchmarks — e chega à verdade sem glamour de que uma ferramenta que pareceu segura e ajudou o processo não ajudou demonstravelmente pacientes em duas semanas, e que detectar qualquer benefício desse tipo 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

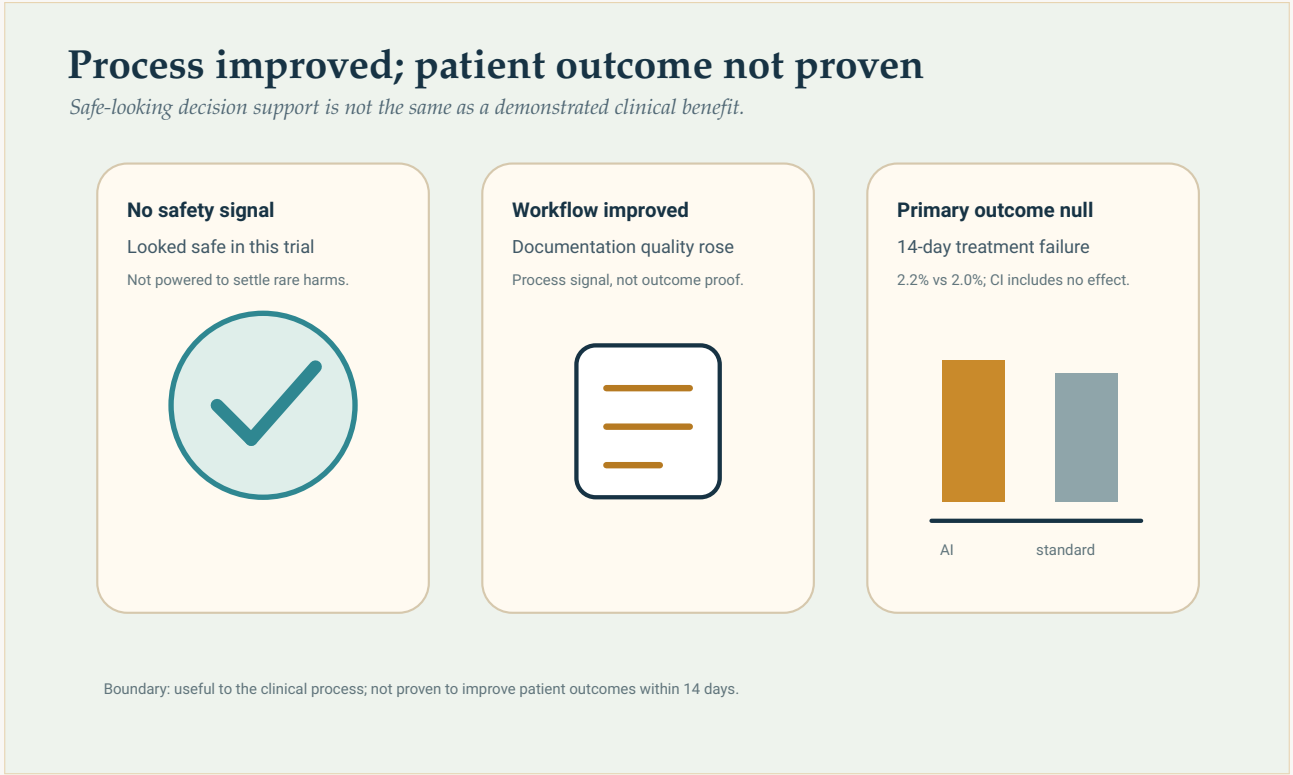
Seguro e útil não é o mesmo que eficaz

A maioria das manchetes sobre IA médica é construída a partir do tipo errado de estudo. Um modelo vai bem em perguntas de prova, ou vence médicos em vinhetas selecionadas, e a história se escreve sozinha: a máquina está pronta para a clínica.

Este ensaio fez algo mais difícil e mais raro. Colocou uma ferramenta de apoio à decisão com IA generativa na atenção primária real, em unidades reais, com pacientes reais, e fez a pergunta que de fato importa: os pacientes ficaram melhor?

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

Isso não é uma falha do estudo. É o estudo funcionando. É assim que evidência responsável sobre IA em medicina se parece quando é medida em pacientes, em vez de em benchmarks.



A ferramenta não mostrou sinal de segurança e melhorou a documentação clínica, mas o desfecho prespecificado de pacientes em 14 dias não melhorou significativamente. Ajuda de processo não é o mesmo que benefício comprovado ao paciente.

O que os autores fizeram

A equipe conduziu um ensaio pragmático, randomizado por conglomerados, em **16 unidades de atenção primária** operadas por uma rede privada de saúde (Penda Health) nos condados de Nairobi e Kiambu, no Quênia. O cuidado nessas unidades é prestado em grande parte por **clinical officers** — profissionais de nível intermediário com diploma de três anos — muitas vezes sem acesso fácil a consulta com profissionais sêniores.

A unidade de randomização foi o clínico, não o paciente. **103 clinical officers** foram randomizados: 52 para o braço de intervenção e 51 para o braço controle. Ambos os braços usaram o mesmo prontuário eletrônico em nuvem. O braço de intervenção teve adicionalmente o **"AI Consult"** (versão 2.0), uma ferramenta de apoio à decisão construída sobre o grande modelo de linguagem **GPT-4o da OpenAI** e incorporada ao prontuário. Ela lia as informações documentadas pelo clínico e podia sinalizar possíveis problemas com o diagnóstico ou o plano de tratamento. Os clínicos mantinham autonomia completa: podiam aceitar, modificar ou ignorar suas sugestões.

> **Qual modelo era, e por que os detalhes importam**
>
> A ferramenta era o **AI Consult 2.0**, rodando **GPT-4o da OpenAI** (a versão de maio de 2025), acessado pela API comercial da OpenAI sob uma licença empresarial e operando em configurações de baixa aleatoriedade (temperatura 0,1). Ela ficava dentro de um prontuário eletrônico sob medida (o EMR da EasyClinic) e era orientada por prompts de sistema escritos para se alinhar às diretrizes nacionais de tratamento do Quênia; os autores publicaram o prompt de instrução completo.
> Por que explicitar isso? Porque o resultado diz respeito a *um sistema específico* — uma versão de modelo, um prompt, um prontuário, um ambiente —, não a "LLMs em medicina" em geral. Os autores fazem o mesmo ponto: chamam seu achado de *benchmark temporal, não uma estimativa fixa de capacidade*. Um modelo mais novo, um prompt diferente ou uma clínica menos digitalizada poderiam todos mover o resultado.
> Sobre independência: a OpenAI depois forneceu apoio em espécie (créditos de computação em nuvem e orientação técnica sobre o uso de sua API), mas os autores afirmam que a decisão de usar OpenAI foi tomada *antes* dessa oferta, e que a OpenAI não teve papel no desenho do ensaio, na coleta de dados, na análise ou na decisão de publicar.

Entre 22 de abril e 16 de julho de 2025, **9.691 pacientes** foram incluídos. O **desfecho primário era deliberadamente centrado no paciente e estrito**: um *composto de falha de tratamento* adjudicado por especialistas dentro de 14 dias da consulta — um painel de clínicos julgou, cego ao braço do estudo, se cada paciente teve um desfecho ruim, como doença não resolvida ou piora. O ensaio foi registrado previamente (Pan-African Clinical Trials Registry 202502499779176).

Essa escolha de desenho é o ponto. É fácil mostrar que uma ferramenta de IA muda o que um clínico escreve. É muito mais difícil, e muito mais significativo, mostrar que ela muda o que acontece com o paciente.

O que encontraram

O desfecho primário não melhorou. Falha de tratamento ocorreu em 102 de 4.693 pacientes (2,2%) no braço IA e 94 de 4.654 (2,0%) no braço controle. As porcentagens brutas foram fracionalmente mais altas com IA, mas depois do ajuste a estimativa pontual inclinou-se para benefício: a **razão de chances ajustada foi 0,77 (intervalo de confiança de 95% 0,55 a 1,08, P = 0,13)** — não estatisticamente significativa. Essa inversão entre números brutos e ajustados não é um deslize aritmético; o ajuste leva em conta diferenças entre os conglomerados de clínicos. De qualquer modo, o intervalo de confiança inclui confortavelmente "sem efeito", então nenhum benefício pode ser alegado, e em termos absolutos o efeito foi minúsculo.

Para uma leitura em linguagem simples de razões de chances, intervalos de confiança e valores de P em conjunto, veja o guia para resultados clínicos.

Sem sinal de segurança — dentro de limites. Nenhum evento adverso grave foi julgado relacionado à ferramenta, e uma revisão independente não encontrou sinal de segurança. Os autores são honestos sobre o teto dessa tranquilização: o ensaio não tinha poder para detectar danos graves raros, e não tinha uma estrutura formal prespecificada de não inferioridade ou segurança, então não pode *provar* segurança para eventos incomuns.

A documentação melhorou. Entre 2.000 atendimentos revisados por especialistas cegos, clínicos usando AI Consult produziram documentação clínica melhor em todos os domínios avaliados — o diagnóstico registrado, o plano de tratamento e a completude geral.

A prescrição quase não se moveu. Não houve diferença significativa em prescrição, incluindo uso correto de antibióticos (razão de chances ajustada 0,86, IC 95% 0,48 a 1,55). A ferramenta não mudou as taxas de prescrição de antibióticos.

Os pacientes não notaram diferença. Entre 826 pacientes que responderam a uma pesquisa de satisfação, a satisfação foi essencialmente idêntica entre os braços, e os tempos de consulta foram semelhantes.

Os custos apontaram ligeiramente para baixo. Em uma análise ajustada, os custos relacionados a antibióticos foram menores no braço IA — plausivelmente por escolhas mais baratas, não por menos prescrições — e a economia por paciente em antibióticos pareceu exceder o custo por paciente de rodar a ferramenta. Os autores sinalizam isso como sugestivo, não resolvido: uma avaliação completa do custo total de propriedade ficou fora do escopo do ensaio.

O resumo em uma linha dos próprios autores é a formulação mais precisa: a assistência por LLM foi segura dentro desses limites, mas não reduziu falha de tratamento em 14 dias, e qualquer benefício provavelmente é modesto.

Por que “sem diferença significativa” não é “não funciona”

É fácil extrapolar demais um desfecho primário nulo em qualquer direção. Duas coisas impedem a história simples.

Primeiro, **o ensaio foi construído para captar um efeito maior do que encontrou.** Desfechos ruins graves na atenção primária são raros — cerca de 2% aqui —, então distinguir um pequeno benefício real de ruído exige números enormes. Os cálculos de poder post hoc dos próprios autores sugerem que detectar um efeito do tamanho observado exigiria um **ensaio muito maior, da ordem de mais de 100.000 pacientes.** Um resultado não significativo em um estudo deste tamanho não descarta um pequeno benefício real; significa que este estudo não conseguiu resolvê-lo.

Segundo, **a comparação foi parcialmente borrada.** Um erro de configuração deu brevemente a alguns clínicos do braço controle acesso ao AI Consult, e clínicos de uma rede compartilhada conversam entre si e carregam hábitos através da fronteira. Ambos os efeitos tendem a tornar os dois braços mais parecidos, empurrando qualquer diferença real para zero. Além disso, a rede anfitriã já operava com padrões relativamente altos, o que deixa menos espaço para uma ferramenta mostrar melhora.

Nada disso resgata uma manchete de “avanço”. Mas significa que a leitura correta é equilibrada, não simplesmente negativa: no desfecho mais exigente e mais informativo, esta ferramenta não ajudou demonstravelmente pacientes em duas semanas — embora tenha ajudado de forma mensurável a documentação e parecido segura.

O que isso *não* prova

- Não mostra que a IA melhorou desfechos de pacientes. No desfecho primário de 14 dias, não houve benefício significativo.
- Não mostra que a IA é inútil. A estimativa pontual favoreceu a ferramenta, a documentação melhorou de ponta a ponta e os custos de medicamentos tenderam a cair; o resultado nulo é compatível com um pequeno benefício real que o ensaio foi pequeno demais para confirmar.
- Não prova que a ferramenta é segura para danos raros. Ela não mostrou sinal de segurança, mas o ensaio não tinha poder nem desenho para certificar segurança para eventos graves incomuns.
- Não mostra que “IA vence médicos” ou substitui clínicos. Isto é *apoio* à decisão; o clínico manteve autoridade completa para aceitar ou rejeitar.
- Não generaliza automaticamente. O ensaio ocorreu em uma única rede privada urbana no Quênia; ambientes rurais, periurbanos e de renda mais alta poderiam diferir em qualquer direção.
- Não estabelece economia de custos. O sinal de custo é sugestivo, não uma avaliação econômica completa.

Quão sólidas são as evidências?

Para a alegação central — nenhuma redução comprovada em dano de curto prazo ao paciente — a evidência é forte *como desenho*, e apropriadamente humilde *como conclusão*. Um ensaio prospectivo, pré-registrado, randomizado por conglomerados, com desfecho composto em nível de paciente e avaliação cega, é próximo da melhor evidência de mundo real que se pode reunir para uma ferramenta assim. É muito mais informativo que uma pontuação de benchmark ou um estudo de vinhetas.

Para os achados secundários — documentação melhor, prescrição inalterada, satisfação semelhante, custos menores com antibióticos — a evidência é boa, mas deve ser lida como secundária: sinais de apoio, não a manchete, e vulneráveis aos mesmos limites de contaminação e rede única.

Para segurança, a evidência é tranquilizadora, mas delimitada: nenhum sinal encontrado, mas não um estudo construído para encontrar danos raros.

A postura mais útil não é nem "funciona" nem "falhou". É: um ensaio cuidadoso no mundo real encontrou que esta ferramenta de IA não levantou sinal de segurança e melhorou o processo de cuidado, sem demonstrar benefício em desfechos de pacientes em duas semanas — e que detectar qualquer benefício desse tipo exigiria um estudo muito maior.

Por que isso importa

O debate sobre IA médica carece exatamente desse tipo de evidência. Há milhares de artigos mostrando modelos acertando provas e igualando clínicos em casos arrumados. Há pouquíssimos ensaios grandes, pragmáticos e randomizados medindo se pacientes reais ficam melhor. Este é um deles, e chega à verdade sem glamour: passar no teste não é o mesmo que ajudar o paciente.

Essa lacuna é a história inteira. Uma ferramenta pode ser genuinamente útil para clínicos — notas mais claras, contas de medicamentos menores, um segundo par de olhos — e ainda assim não mover um desfecho duro de paciente em quinze dias. Os dois fatos podem ser verdadeiros ao mesmo tempo, e um sistema de saúde maduro precisa segurá-los juntos em vez de escolher o conveniente.

Também reajusta o ônus da prova em uma direção útil. Se uma empresa quer afirmar que sua IA clínica melhora o cuidado, a evidência relevante não é um ranking. É um ensaio como este, em desfechos que importam aos pacientes — e, idealmente, um maior, porque a lição honesta aqui é que benefícios modestos precisam de estudos grandes para aparecer.

Em resumo

Um ensaio pragmático, randomizado por conglomerados, em 16 unidades de atenção primária no Quênia testou uma ferramenta de apoio à decisão com IA generativa ("AI Consult") adicionada ao prontuário eletrônico usado por clinical officers. Entre 9.691 pacientes, o composto de falha de trata-

mento adjudicado por especialistas dentro de 14 dias foi 2,2% com a ferramenta versus 2,0% sem ela (razão de chances ajustada 0,77, IC 95% 0,55–1,08, P = 0,13) — sem diferença significativa. A ferramenta não mostrou sinal de segurança, melhorou a documentação clínica em todos os domínios avaliados, não mudou a prescrição, deixou a satisfação dos pacientes inalterada e foi associada a custos de antibióticos um pouco menores. Os autores concluem que ela foi segura dentro desses limites, mas não reduziu falha de tratamento, com qualquer benefício provavelmente modesto; detectar um efeito do tamanho observado exigiria um ensaio muito maior (da ordem de 100.000 pacientes). O resultado não mostra que IA clínica melhora desfechos de pacientes, nem que é inútil — mostra que uma ferramenta que pareceu segura e ajudou o processo não ajudou demonstravelmente pacientes em duas semanas, em uma rede privada urbana.

Leitura crítica

O que o artigo mostra: Em um ensaio randomizado no mundo real, adicionar uma ferramenta de apoio à decisão baseada em LLM aos prontuários da atenção primária não levantou sinal de segurança, melhorou a qualidade da documentação clínica, não mudou a prescrição e não reduziu significativamente um composto estrito de falha de tratamento em pacientes em 14 dias.

O que é plausível, mas não provado: Que a ferramenta produza uma pequena redução real em falha de tratamento, pequena demais para este ensaio detectar; que economize dinheiro quando todos os custos forem contados; que documentação melhor eventualmente se traduza em melhor cuidado.

O que ele não mostra: Que IA clínica melhore desfechos de pacientes; que seja insegura para eventos raros; que substitua ou supere clínicos; que estes resultados se transfiram a ambientes rurais ou de renda mais alta; que a economia de custos esteja estabelecida.

Principais limitações: Poder estatístico para um efeito maior que o observado (desfechos raros exigem amostras muito grandes); um erro de configuração contaminou o braço controle e hábitos compartilhados na rede borram a comparação, ambos enviesando para nenhuma diferença; uma única rede privada urbana com padrões já altos; sem estrutura prespecificada de não inferioridade ou segurança; horizonte curto de 14 dias.

Quanta confiança um leitor geral deve ter? Alta de que a ferramenta não mostrou sinal de segurança neste ensaio e melhorou a documentação. Alta de que não melhorou demonstravelmente desfechos de pacientes em 14 dias aqui. Baixa para qualquer alegação de que ela “funciona” ou “falha” como intervenção de benefício ao paciente — essa pergunta está genuinamente sem resolução e precisa de um estudo muito maior. Postura apropriada: um resultado cuidadoso de mundo real sobre uma ferramenta que não levantou sinal de segurança e melhorou o processo de cuidado, não um veredito de que a IA transforma — ou arruína — a atenção primária.

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