The Clean Paper

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

AI klinikal yang kelihatan selamat dan memperbaik dokumentasi — tetapi tidak meningkatkan hasil pesakit

Ujian pragmatik rawak berkelompok di 16 fasiliti penjagaan primer Kenya menguji alat sokongan keputusan AI generatif yang ditambahkan pada rekod yang digunakan oleh pegawai klinikal. Dalam kalangan 9,691 pesakit, hasil gabungan kegagalan rawatan 14 hari yang ketat dan dinilai pakar ialah 2.2% dengan alat berbanding 2.0% tanpa alat (nisbah odds terlaras 0.77, 95% CI 0.55-1.08, P = 0.13) — tiada perbezaan signifikan. Alat itu tidak menunjukkan isyarat keselamatan, meningkatkan dokumentasi dalam semua domain yang dinilai, tidak mengubah preskripsi atau kepuasan, dan cenderung menurunkan kos antibiotik. Itu bukan kajian yang gagal; ia kajian yang jarang dan jujur — diukur pada pesakit, bukan penanda aras — dan hasilnya ialah kebenaran yang tidak glamor bahawa alat yang kelihatan selamat dan membantu proses tidak terbukti membantu pesakit dalam dua minggu, dan untuk mengesan manfaat sedemikian memerlukan ujian jauh lebih besar.

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

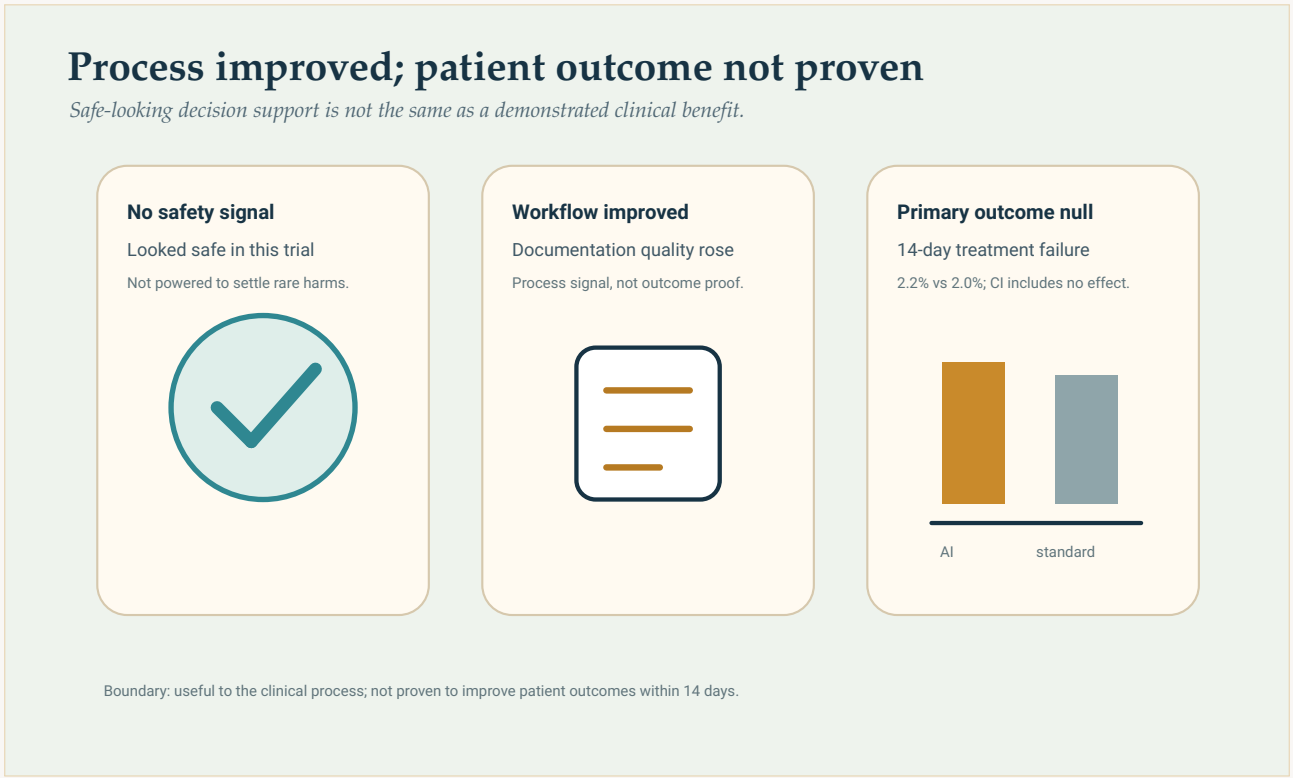
Selamat dan membantu tidak sama dengan berkesan

Kebanyakan tajuk berita tentang AI perubatan dibina daripada jenis kajian yang salah. Sebuah model mendapat markah tinggi dalam soalan peperiksaan, atau mengatasi doktor pada vignette yang dipilih rapi, lalu ceritanya seolah-olah sudah siap: mesin itu sudah bersedia untuk klinik.

Ujian ini melakukan sesuatu yang lebih sukar dan lebih jarang. Ia meletakkan alat sokongan keputusan berasaskan AI generatif dalam penjagaan primer sebenar, di fasiliti sebenar, dengan pesakit sebenar, lalu menanyakan soalan yang benar-benar penting: adakah keadaan pesakit menjadi lebih baik?

Jawapan yang jujur ialah tidak — sekurang-kurangnya tidak secara terukur dalam 14 hari. Alat itu tidak menunjukkan isyarat keselamatan yang membimbangkan. Ia meningkatkan kualiti dokumentasi klinikal. Ia mungkin malah mengurangkan sebahagian kos ubat. Tetapi ia tidak mengurangkan kegagalan rawatan secara signifikan, dan para penulis berhati-hati menyatakan bahawa sebarang manfaat, jika memang ada, mungkin sederhana.

Itu bukan kegagalan kajian. Itulah kajian yang berfungsi sebagaimana sepatutnya. Beginilah rupa bukti yang bertanggungjawab tentang AI dalam perubatan apabila ia diukur pada pesakit, bukannya pada penanda aras.



Alat itu tidak menunjukkan isyarat keselamatan yang membimbangkan dan meningkatkan dokumentasi klinikal, tetapi hasil pesakit 14 hari yang ditetapkan lebih awal tidak bertambah baik secara signifikan. Membantu proses tidak sama dengan manfaat kepada pesakit yang sudah terbukti.

Apa yang dilakukan oleh para penulis

Pasukan itu menjalankan ujian pragmatik rawak berkelompok di **16 fasiliti penjagaan primer** yang dikendalikan oleh rangkaian kesihatan swasta (Penda Health) di daerah Nairobi dan Kiambu, Kenya. Penjagaan di fasiliti ini sebahagian besarnya diberikan oleh **pegawai perubatan** — pengamal peringkat pertengahan dengan diploma tiga tahun — yang sering tidak mempunyai akses mudah kepada konsultasi kanan.

Unit pengacakan ialah klinisian, bukan pesakit. **103 pegawai perubatan** dirawakkan: 52 ke kumpulan intervensi dan 51 ke kumpulan kawalan. Kedua-dua kumpulan menggunakan sistem rekod perubatan elektronik berasaskan awan yang sama. Kumpulan intervensi turut mempunyai **"AI Consult"** (versi 2.0), alat sokongan keputusan yang dibina berasaskan model bahasa besar **GPT-4o OpenAI** dan disepadukan dalam rekod itu. Ia membaca maklumat yang didokumentasikan oleh klinisian dan boleh menandakan kemungkinan masalah pada diagnosis atau pelan rawatan. Klinisian kekal mempunyai autonomi penuh: mereka boleh menerima, mengubah atau mengabaikan cadangannya.

> **Model apakah yang digunakan, dan mengapa butirannya penting**
>
> Alat itu ialah **AI Consult 2.0**, yang menggunakan **GPT-4o OpenAI** (keluaran Mei 2025), diakses melalui API komersial OpenAI di bawah lesen perusahaan dan dijalankan pada tetapan rawak yang rendah (suhu 0.1). Ia berada di dalam rekod elektronik yang dibina khusus (EMR EasyClinic) dan dipandu oleh prompt sistem yang ditulis agar selaras dengan garis panduan rawatan kebangsaan Kenya; para penulis menerbitkan prompt arahan penuh.
> Mengapa perlu dinyatakan sedemikian terperinci? Kerana hasilnya berkaitan dengan *satu sistem tertentu* — satu versi model, satu prompt, satu rekod, satu tetapan — bukannya tentang "LLM dalam perubatan" secara umum. Para penulis menegaskan perkara yang sama: mereka menyebut penemuan ini sebagai *penanda aras pada suatu masa, bukannya anggaran keupayaan yang tetap*. Model yang lebih baharu, prompt yang berbeza atau klinik yang kurang terdigital semuanya boleh mengubah hasil.
> Tentang kebebasan kajian: OpenAI kemudian memberikan sokongan dalam bentuk bukan tunai (kredit pengkomputeran awan dan panduan teknikal untuk penggunaan APInya), tetapi para penulis menyatakan bahawa keputusan untuk menggunakan OpenAI dibuat *sebelum* tawaran itu, dan OpenAI tidak memainkan peranan dalam reka bentuk ujian, pengumpulan data, analisis atau keputusan untuk menerbitkan hasil.

Antara 22 April dan 16 Julai 2025, **9,691 pesakit** didaftarkan. **Hasil utama sengaja ditetapkan sebagai ukuran yang ketat dan berpusat pada pesakit**: satu *hasil gabungan kegagalan rawatan* dalam tempoh 14 hari selepas lawatan yang diadili oleh pakar — panel klinisian yang tidak mengetahui kumpulan kajian menilai sama ada setiap pesakit mengalami hasil buruk seperti penyakit yang tidak pulih atau bertambah teruk. Ujian ini didaftarkan lebih awal (Pan-African Clinical Trials Registry 202502499779176).

Pilihan reka bentuk itu ialah perkara utamanya. Mudah untuk menunjukkan bahawa alat AI mengubah apa yang ditulis oleh klinisian. Jauh lebih sukar, dan jauh lebih bermakna, untuk menunjukkan bahawa ia mengubah apa yang benar-benar berlaku kepada pesakit.

Apa yang mereka temui

Hasil utama tidak bertambah baik. Kegagalan rawatan berlaku pada 102 daripada 4,693 pesakit (2.2%) dalam kumpulan AI dan 94 daripada 4,654 (2.0%) dalam kumpulan kawalan. Peratusan mentah sedikit lebih tinggi dalam kumpulan AI, tetapi selepas pelarasan, anggaran titik bergerak ke arah kemungkinan manfaat: **nisbah odds terlaras ialah 0.77 (selang keyakinan 95% 0.55 hingga 1.08, P = 0.13)** — tidak signifikan secara statistik. Perubahan arah antara angka mentah dan angka terlaras itu bukan kesilapan pengiraan; pelarasan mengambil kira perbezaan antara kelompok klinisian. Dalam kedua-dua cara membaca data, selang keyakinan jelas merangkumi "tiada kesan", jadi manfaat tidak boleh didakwa, dan secara mutlak kesannya amat kecil.

Untuk cara mudah membaca nisbah odds, selang keyakinan dan nilai P bersama-sama, lihat panduan membaca hasil klinikal.

Tiada isyarat keselamatan — dalam batas tertentu. Tiada kejadian buruk serius dinilai berkaitan dengan alat itu, dan semakan bebas tidak menemui isyarat keselamatan. Para penulis jujur tentang had jaminan ini: ujian tersebut tidak mempunyai kuasa statistik untuk mengesan kemudaratan serius yang jarang berlaku, dan tidak mempunyai rangka kerja noninferiority atau keselamatan formal yang ditetapkan lebih awal, jadi ia tidak boleh *membuktikan* keselamatan untuk kejadian luar biasa.

Dokumentasi menjadi lebih baik. Dalam 2,000 pertemuan klinikal yang disemak oleh pakar yang dibutakan kepada kumpulan, klinisian yang menggunakan AI Consult menghasilkan dokumentasi klinikal lebih baik dalam semua domain yang dinilai — diagnosis yang direkodkan, pelan rawatan dan kelengkapan keseluruhan.

Corak preskripsi hampir tidak berubah. Tiada perbezaan signifikan dalam preskripsi, termasuk penggunaan antibiotik yang betul (nisbah odds terlaras 0.86, 95% CI 0.48 hingga 1.55). Alat itu tidak mengubah kadar preskripsi antibiotik.

Pesakit tidak menyedari perbezaan. Dalam kalangan 826 pesakit yang melengkapkan tinjauan kepuasan, kepuasan pada asasnya sama antara kedua-dua kumpulan, dan tempoh konsultasi juga serupa.

Kos menunjukkan sedikit penurunan. Dalam analisis terlaras, kos berkaitan antibiotik lebih rendah dalam kumpulan AI — mungkin melalui pilihan yang lebih murah, bukannya preskripsi yang lebih sedikit — dan penjimatan antibiotik bagi setiap pesakit kelihatan melebihi kos menjalankan alat itu bagi setiap pesakit. Para penulis menandakan hasil ini sebagai petunjuk, bukan kesimpulan: perakaunan penuh jumlah kos pemilikan berada di luar skop ujian.

Ringkasan satu ayat daripada penulis sendiri ialah rumusan yang paling jelas: bantuan LLM selamat dalam batas tersebut tetapi tidak mengurangkan kegagalan rawatan dalam 14 hari, dan sebarang manfaat mungkin sederhana.

Mengapa “tiada perbezaan signifikan” bukan bermaksud “ia tidak berfungsi”

Hasil utama yang null mudah ditafsir secara berlebihan ke mana-mana arah. Dua perkara menghalang cerita yang terlalu ringkas.

Pertama, **ujian ini direka untuk menangkap kesan yang lebih besar daripada yang ditemui.** Hasil buruk yang serius dalam penjagaan primer jarang berlaku — sekitar 2% di sini — jadi membezakan manfaat sebenar yang kecil daripada hingar memerlukan jumlah peserta yang sangat besar. Pengiraan kuasa selepas kajian para penulis sendiri mencadangkan bahawa untuk mengesan kesan sebesar yang mereka perhatikan memerlukan **ujian yang jauh lebih besar, pada skala lebih daripada 100,000 pesakit.** Hasil tidak signifikan dalam kajian sebesar ini tidak menolak manfaat sebenar yang kecil; ia bermaksud kajian ini tidak mampu membezakannya dengan jelas.

Kedua, **perbandingan antara kumpulan menjadi sedikit kabur.** Kesilapan konfigurasi memberi sebilangan klinisian kumpulan kawalan akses sementara kepada AI Consult, dan klinisian dalam rangkaian yang sama bercakap sesama sendiri serta membawa tabiat merentasi sempadan kumpulan. Kedua-dua kesan ini cenderung menjadikan dua kumpulan lebih serupa, lalu menolak perbezaan sebenar ke arah sifar. Selain itu, rangkaian yang menjadi tuan rumah sudah sudah beroperasi pada tahap piawai yang agak tinggi, meninggalkan ruang yang lebih kecil untuk alat menunjukkan peningkatan.

Tiada satu pun daripada ini menyelamatkan tajuk “kejayaan besar”. Tetapi ia bermaksud tafsiran yang betul perlu berhati-hati, bukan sekadar mengecilkan hasil: pada titik akhir yang paling sukar dan paling jujur, alat ini tidak terbukti membantu pesakit dalam dua minggu — walaupun ia memang membantu penyimpanan rekod secara terukur dan kelihatan selamat.

Apa yang tidak dibuktikan oleh kajian ini

- Ia **tidak** menunjukkan bahawa AI meningkatkan hasil pesakit. Pada titik akhir utama 14 hari, tiada manfaat signifikan.
- Ia **tidak** menunjukkan bahawa AI tidak berguna. Anggaran titik memihak kepadanya, dokumentasi bertambah baik secara menyeluruh, dan kos ubat cenderung turun; hasil null masih serasi dengan manfaat sebenar yang kecil tetapi terlalu kecil untuk disahkan oleh ujian ini.
- Ia **tidak** membuktikan alat itu selamat daripada kemudaratan yang jarang berlaku. Tiada isyarat keselamatan ditemui, tetapi kajian ini tidak mempunyai kuasa atau reka bentuk untuk mengesahkan keselamatan bagi kejadian serius yang jarang berlaku.
- Ia **tidak** menunjukkan bahawa “AI mengalahkan doktor” atau menggantikan klinisian. Ini ialah *sokongan* keputusan; klinisian kekal mempunyai kuasa penuh untuk menerima atau menolaknya.
- Ia **tidak** boleh digeneralisasikan secara automatik. Ujian berlaku dalam satu rangkaian swasta bandar di Kenya; persekitaran luar bandar, pinggir bandar atau negara berpendapatan tinggi boleh berbeza ke mana-mana arah.
- Ia **tidak** membuktikan penjimatan kos. Isyarat kos itu memberi petunjuk, bukan penilaian

ekonomi yang lengkap.

Sejauh mana kukuh buktinya?

Bagi dakwaan utama — tiada pengurangan yang terbukti dalam kemudaratan jangka pendek kepada pesakit — buktinya kukuh *dari segi reka bentuk*, dan sewajarnya sederhana *dari segi kesimpulan*. Ujian prospektif, pradaftar, rawak berkelompok dengan hasil gabungan pada peringkat pesakit yang dinilai secara buta hampir kepada bukti dunia sebenar terbaik yang boleh dikumpulkan untuk alat seperti ini. Ia jauh lebih bermaklumat daripada skor penanda aras atau kajian vignette.

Bagi penemuan sekunder — dokumentasi lebih baik, preskripsi tidak berubah, kepuasan serupa, kos antibiotik lebih rendah — buktinya baik tetapi harus dibaca sebagai sekunder: isyarat sokongan, bukan tajuk utama, dan terdedah kepada batas pencemaran antara kumpulan serta hanya satu rangkaian.

Bagi keselamatan, bukti itu meyakinkan tetapi terbatas: tiada isyarat ditemui, namun ini bukan kajian yang dibina untuk mencari kemudaratan jarang berlaku.

Sikap yang paling berguna bukanlah “ia berfungsi” mahupun “ia gagal”. Sebaliknya: satu ujian dunia sebenar yang teliti mendapati alat AI ini tidak menimbulkan isyarat keselamatan dan memperbaik proses penjagaan, tanpa menunjukkan manfaat pada hasil pesakit dalam dua minggu — dan untuk mengesan manfaat sedemikian mungkin memerlukan kajian yang jauh lebih besar.

Mengapa ia penting

Perdebatan tentang AI perubatan sangat kekurangan bukti seperti ini. Terdapat ribuan makalah yang menunjukkan model cemerlang dalam peperiksaan dan menyamai klinisian pada kes yang kemas. Tetapi sangat sedikit ujian pragmatik rawak berskala besar yang mengukur sama ada pesakit sebenar benar-benar mendapat manfaat. Ini salah satunya, dan hasilnya membawa kita kepada kebenaran yang tidak glamor: lulus ujian tidak sama dengan membantu pesakit.

Jurang itulah keseluruhan ceritanya. Sebuah alat boleh benar-benar berguna kepada klinisian — nota lebih jelas, bil ubat lebih rendah, sepasang mata tambahan — namun masih tidak mengubah hasil pesakit yang keras dalam dua minggu. Kedua-duanya boleh benar pada masa yang sama, dan sistem kesihatan yang matang perlu memegang kedua-dua fakta itu bersama-sama, bukan memilih yang paling mudah.

Ia juga menetapkan semula beban pembuktian ke arah yang berguna. Jika sebuah syarikat mahu mendakwa AI klinikalnya meningkatkan penjagaan, bukti yang relevan bukan papan kedudukan. Ia ialah ujian seperti ini, pada hasil yang penting kepada pesakit — dan sebaik-baiknya ujian lebih besar, kerana pelajaran jujur di sini ialah manfaat sederhana memerlukan kajian besar untuk dilihat.

Ringkasan utama

Ujian pragmatik rawak berkelompok di 16 fasiliti penjagaan primer Kenya menguji alat sokongan keputusan AI generatif ("AI Consult") yang ditambahkan pada rekod elektronik yang digunakan oleh pegawai perubatan. Dalam kalangan 9,691 pesakit, hasil gabungan kegagalan rawatan dalam 14 hari yang dinilai pakar ialah 2.2% dengan alat berbanding 2.0% tanpa alat (nisbah odds terlaras 0.77, 95% CI 0.55–1.08, P = 0.13) — tiada perbezaan signifikan. Alat itu tidak menunjukkan isyarat keselamatan, meningkatkan dokumentasi klinikal dalam semua domain yang dinilai, tidak mengubah preskripsi, tidak mengubah kepuasan pesakit, dan dikaitkan dengan kos antibiotik yang agak lebih rendah. Para penulis menyimpulkan bahawa ia selamat dalam batas tersebut tetapi tidak mengurangkan kegagalan rawatan, dengan sebarang manfaat mungkin sederhana; untuk mengesan kesan sebesar yang diperhatikan memerlukan ujian jauh lebih besar (pada skala 100,000 pesakit). Hasil ini tidak menunjukkan bahawa AI klinikal meningkatkan hasil pesakit, dan juga tidak menunjukkan bahawa ia tidak berguna — ia menunjukkan bahawa alat yang kelihatan selamat dan membantu proses tidak terbukti membantu pesakit dalam dua minggu, dalam satu rangkaian swasta bandar.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh makalah: Dalam ujian rawak dunia sebenar, menambah alat sokongan keputusan berasaskan LLM kepada rekod penjagaan primer tidak menimbulkan isyarat keselamatan, meningkatkan kualiti dokumentasi klinikal, tidak mengubah preskripsi, dan tidak mengurangkan secara signifikan hasil gabungan kegagalan rawatan pesakit yang ketat dalam 14 hari.

Apa yang munasabah tetapi belum dibuktikan: Bahawa alat itu menghasilkan pengurangan kecil sebenar dalam kegagalan rawatan yang terlalu kecil untuk dikesan oleh ujian ini; bahawa ia menjimatkan wang apabila semua kos dikira; bahawa dokumentasi lebih baik akhirnya membawa kepada penjagaan lebih baik.

Apa yang tidak ditunjukkan: Bahawa AI klinikal meningkatkan hasil pesakit; bahawa ia selamat daripada kejadian jarang; bahawa ia menggantikan atau mengatasi klinisian; bahawa hasil ini boleh dipindahkan terus ke kawasan luar bandar atau negara berpendapatan tinggi; bahawa penjimatan kos sudah terbukti.

Had utama: Kuasa statistik ditetapkan untuk kesan yang lebih besar daripada yang diperhatikan (hasil jarang memerlukan sampel sangat besar); kesilapan konfigurasi mencemari kumpulan kawalan dan tabiat dalam rangkaian yang sama mengaburkan perbandingan, kedua-duanya cenderung menolak perbezaan ke arah sifar; satu rangkaian swasta bandar yang sudah mempunyai piawai tinggi; tiada rangka kerja noninferiority atau keselamatan yang ditetapkan lebih awal; tempoh hanya 14 hari.

Berapa banyak keyakinan patut diberikan pembaca umum? Tinggi bahawa alat itu tidak menunjukkan isyarat keselamatan dalam ujian ini dan meningkatkan dokumentasi. Tinggi bahawa ia tidak terbukti meningkatkan hasil pesakit 14 hari di sini. Rendah bagi sebarang dakwaan bahawa ia "berfungsi" atau "gagal" sebagai intervensi manfaat pesakit — soalan itu benar-benar belum

selesai dan memerlukan kajian jauh lebih besar. Sikap yang wajar: hasil dunia sebenar yang teliti tentang alat yang tidak menimbulkan isyarat keselamatan dan memperbaik proses penjagaan, bukan keputusan bahawa AI mengubah — atau merosakkan — penjagaan primer.

Sumber

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