



The CLEAN PAPER

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

هوش مصنوعی در مراقبت اولیه مستندسازی را بهتر کرد، اما بیماران ظرف ۱۴ روز بهتر نشدند

یک کارآزمایی عمل گرایانه و خوشه‌ای-تصادفی در ۱۶ مرکز مراقبت اولیه کینیا، ابزار *Consult AI* مبتنی بر *GPT-4o* را در پرونده الکترونیک ۱۰۳ کارشناس بالینی آزمود. میان ۹۶۹۱ بیمار، شکست درمان دآوری شده در ۱۴ روز ۲٫۲٪ با *AI* و ۲٫۰٪ بدون آن بود؛ *OR* تعدیل شده ۰٫۷۷ با فاصله اطمینان ۹۵٪ برابر ۰٫۵۵-۱٫۰۸ و $P=0.13$ ، بنابراین سود معناداری نشان داده نشد. ابزار سیگنال ایمنی تازه‌ای ایجاد نکرد، مستندسازی را بهتر کرد و شاید هزینه برخی داروها را کاهش داده باشد. نتیجه نمونه خوبی است از اینکه بهبود فرایند بالینی با بهبود اثبات شده پیامد بیماریکی نیست.

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: 6-04503-026-s41591/1038.10

ایمن و مفید بودن با «موثر بودن» یکی نیست

بیشتر تیرهای مربوط به هوش مصنوعی پزشکی از نوع اشتباهی از مطالعه ساخته می‌شوند. یک مدل در سؤال‌های امتحانی نمره خوبی می‌گیرد یا در های vignette گزینش شده از پزشکان جلو می‌زند و داستان خودش نوشته می‌شود: ماشین برای کلینیک آماده است.

این کارآزمایی کاری دشوارتر و نادرتر انجام داد. ابزار پشتیبان تصمیم‌گیری مبتنی بر هوش مصنوعی مولد را وارد مراقبت اولیه واقعی، در مراکز واقعی و کنار بیماران واقعی کرد و پرسشی را مطرح کرد که واقعاً اهمیت دارد: آیا حال بیماران بهتر شد؟

پاسخ صادقانه نه است — دست کم نه به شکلی که در ۱۴ روز قابل اندازه‌گیری باشد. ابزار هیچ سیگنال ایمنی نشان نداد. کیفیت مستندسازی بالینی را بهتر کرد. شاید حتی هزینه‌های دارویی را پایین آورده باشد. اما شکست درمان را به‌طور معنادار کاهش نداد و نویسندگان با دقت می‌گویند اگر منفعتی وجود داشته باشد، احتمالاً متوسط است.

این شکست مطالعه نیست. این یعنی مطالعه درست کار کرده. وقتی هوش مصنوعی پزشکی را به جای benchmark روی بیمار اندازه بگیریم، شواهد مسئولانه شبیه همین‌اند.

Process improved; patient outcome not proven

Safe-looking decision support is not the same as a demonstrated clinical benefit.

No safety signal

Looked safe in this trial

Not powered to settle rare harms.



Workflow improved

Documentation quality rose

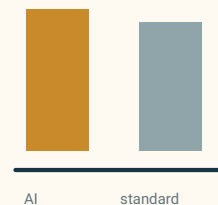
Process signal, not outcome proof.



Primary outcome null

14-day treatment failure

2.2% vs 2.0%; CI includes no effect.



Boundary: useful to the clinical process; not proven to improve patient outcomes within 14 days.

ابزار هیچ سیگنال ایمنی نشان نداد و مستندسازی بالینی را بهتر کرد، اما پیامد از پیش تعیین شده ۱۴ روزه بیمار به‌طور معنادار بهتر نشد. کمک به فرایند مراقبت با منفعت اثبات شده برای بیمار یکی نیست.

نویسندگان چه کردند

گروه یک کارآزمایی pragmatic و cluster-randomized در ۱۶ مرکز مراقبت اولیه متعلق به شبکه خصوصی Health Penda در شهرستان‌های Nairobi و Kiambu کنیا اجرا کرد. مراقبت در این مراکز عمدتاً توسط **clinical officer** – کارکنان بالینی سطح میانی با مدرک سه ساله – ارائه می‌شود که اغلب دسترسی آسانی به مشاوره پزشکی ارشد ندارند.

واحد تصادفی‌سازی پزشکی/کارورز بود، نه بیمار. ۱۰۳ **clinical officer** تصادفی شدند: ۵۲ نفر به بازوی مداخله و ۵۱ نفر به کنترل. هر دو بازو از همان پرونده الکترونیک ابری استفاده کردند. بازوی مداخله افزون بر آن **Consult AI** نسخه 0.2 داشت، ابزار پشتیبان تصمیم‌گیری ساخته شده روی مدل زبانی بزرگ **GPT-4o** از **OpenAI** که در همان پرونده تعبیه شده بود. ابزار اطلاعات ثبت شده توسط **clinician** را می‌خواند و می‌توانست مشکلات احتمالی در تشخیص یا برنامه درمان را علامت بزند. اختیار کامل دست **clinician** بود: می‌توانست پیشنهاد را بپذیرد، تغییر دهد یا نادیده بگیرد.

کدام مدل بود و چرا جزئیات مهم‌اند

مجوز با **OpenAI** تجاری API طریق از را ۲۰۲۵) مه (نسخه **OpenAI** از **GPT-4o** که بود **AI Consult 0.2** ابزار اختصاصی الکترونیک پرونده داخل ابزار می‌کرد. اجرا (1.0 temperature) پایین randomness تنظیم و enterprise بود؛ شده نوشته کنیا درمان ملی دستورالعمل‌های با هم‌راستایی برای هایش **system prompt** و داشت قرار **EasyClinic** کردند. منتشر را **prompt** کامل متن نویسندگان چرا این جزئیات را می‌گوییم؟ چون نتیجه درباره یک سامانه مشخص است – یک نسخه مدل، یک **prompt**، یک پرونده و یک تنظیم – نه درباره **LLMs** «در پزشکی» به‌طور کلی. خود نویسندگان هم همین نکته را می‌گویند و یافته را **benchmark** زمانی، نه برآورد ثابت توانایی می‌نامند. مدل تازه‌تر، **prompt** متفاوت یا کلینیکی با دیجیتالی‌شدن کمتر می‌تواند نتیجه را جابه‌جا کند. درباره استقلال: **OpenAI** بعداً حمایت غیرنقدی فراهم کرد (اعتبار رایانش ابری و راهنمایی فنی برای API)، اما نویسندگان می‌گویند تصمیم استفاده از **OpenAI** پیش از این پیشنهاد گرفته شده بود و **OpenAI** در طراحی کارآزمایی، جمع‌آوری داده، تحلیل یا تصمیم به انتشار نقشی نداشت.

بین ۲۲ آوریل و ۱۶ ژوئیه ۲۰۲۵، ۹۶۹۱ بیمار وارد شدند. پیامد اولیه عمداً بیمارمحور و سخت‌گیرانه بود: یک ترکیب داوری شده توسط متخصص از شکست درمان در ۱۴ روز پس از ویزیت – هیئتی از **clinician** بدون اطلاع از بازوی مطالعه، قضاوت کردند آیا هر بیمار پیامد بدی مانند ادامه یا بدتر شدن بیماری داشته است. کارآزمایی از پیش ثبت شده بود (Pan-African Clinical Registry Trials 202502499779176).

همین انتخاب طراحی اصل ماجراست. آسان است نشان دهید ابزار **AI** چیزی را که **clinician** می‌نویسد تغییر می‌دهد. بسیار سخت‌تر – و مهم‌تر – است نشان دهید آنچه برای بیمار رخ می‌دهد تغییر کرده است.

چه پیدا کردند

پیامد اولیه بهتر نشد. شکست درمان در ۱۰۲ نفر از ۴۶۹۳ بیمار (۲٫۲٪) در بازوی **AI** و ۹۴ نفر از ۴۶۵۴ (۲٫۰٪) در کنترل رخ داد. درصد خام کمی در **AI** بالاتر بود، اما پس از تعدیل نقطه‌برآورد به سوی منفعت متمایل شد: **ratio odds adjusted** برابر ۰٫۷۷ بود (55.0 CI 95% تا 13.0 = P, 08.1) – از نظر آماری معنادار نبود. وارونگی میان عدد خام و تعدیل شده اشتباه حسابی نیست؛ تعدیل اختلاف میان خوشه‌های **clinician** را در نظر می‌گیرد. در هر حال فاصله اطمینان «بدون اثر» را به راحتی شامل می‌کند، پس نمی‌توان ادعای منفعت کرد و اثر مطلق هم بسیار کوچک بود.

برای خوانش ساده odds، ratio، فاصله اطمینان و value P در کنار هم، راهنمای نتایج بالینی را ببینید.

هیچ سیگنال ایمنی — در محدوده این مطالعه. هیچ عارضه جدی ای مرتبط با ابزار تشخیص داده نشد و بازبینی مستقل سیگنال ایمنی پیدا نکرد. نویسندگان سقف این آرامش را صریح می گویند: کارآزمایی برای یافتن آسیب های شدید و نادر توان کافی نداشت و چارچوب از پیش تعیین شده noninferiority یا ایمنی رسمی نداشت، پس نمی تواند ایمنی برای رویدادهای نادر را اثبات کند.

مستندسازی بهتر شد. در میان ۲۰۰۰ ویزیت که متخصصان کور بازبینی کردند، های-clinician استفاده کننده از Consult AI در همه حوزه های نمره گذاری شده مستندسازی بهتری داشتند — تشخیص ثبت شده، برنامه درمان و کامل بودن کلی.

نسخه نویسی تقریباً تکان نخورد. اختلاف معناداری در نسخه نویسی، از جمله مصرف صحیح آنتی بیوتیک، وجود نداشت (adjusted OR 86.0، 95% CI 48.0 تا 55.1). ابزار نرخ نسخه کردن آنتی بیوتیک را تغییر نداد.

بیماران تفاوتی حس نکردند. در ۸۲۶ بیماری که پرسش نامه رضایت را تکمیل کردند، رضایت تقریباً در دو بازوی یکسان و زمان مشاوره مشابه بود.

هزینه ها کمی رو به پایین اشاره کردند. در تحلیل تعدیل شده، هزینه مرتبط با آنتی بیوتیک در بازوی AI کمتر بود — احتمالاً به دلیل انتخاب داروهای ارزان تر، نه نسخه های کمتر — و صرفه جویی آنتی بیوتیک به ازای بیمار ظاهراً از هزینه اجرای ابزار بیشتر بود. نویسندگان این را پیشنهاد کننده می دانند، نه قطعی؛ حسابداری کامل ownership of cost total خارج از کارآزمایی بود.

خلاصه یک خطی خود نویسندگان تمیزترین است: کمک LLM در این محدودیت ها ایمن به نظر رسید، اما شکست درمان در ۱۴ روز را کاهش نداد و هر منفعتی احتمالاً متوسط است.

چرا «تفاوت معنادار نبود» به معنی «کار نمی کند» نیست

پیامد اولیه null را می شود در هر دو جهت بیش از حد خواند. دو نکته جلوی داستان ساده را می گیرند.

اول، کارآزمایی برای اثر بزرگ تری از آنچه دید طراحی شده بود. پیامدهای بد جدی در مراقبت اولیه نادرند — این جا حدود ۲٪ — بنابراین جدا کردن منفعت کوچک واقعی از نویز به تعداد بسیار زیادی بیمار نیاز دارد. محاسبات توان post-hoc نویسندگان نشان می دهد برای تشخیص اثری به اندازه مشاهده شده کارآزمایی بسیار بزرگ تر، در حد بیش از ۱۰۰،۰۰۰ بیمار لازم است. نتیجه غیرمعنادار در این حجم، منفعت کوچک واقعی را رد نمی کند؛ می گوید این مطالعه قادر به تفکیک نبود.

دوم، مقایسه تا حدی آلوده شد. یک خطای پیکربندی برای مدتی به بعضی های-clinician بازوی کنترل دسترسی به Consult AI داد، و افراد در یک شبکه مشترک با هم حرف می زنند و عادت ها را از مرز بازوها عبور می دهند. هر دو عامل دو گروه را شبیه تر می کنند و هر اختلاف واقعی را به سمت صفر هل می دهند. افزون بر آن، شبکه میزبان از ابتدا استاندارد نسبتاً بالایی داشت، پس جای کمتری برای نمایش بهبود باقی می ماند.

هیچ کدام تیتز «جهش بزرگ» را نجات نمی دهد. اما خوانش درست را تنظیم شده نگه می دارد، نه تحقیرآمیز: روی سخت ترین و صادقانه ترین endpoint، این ابزار در دو هفته به طور قابل اثبات به بیماران کمک نکرد — در حالی که مستندسازی را واقعاً بهتر کرد و ایمن به نظر رسید.

این مقاله چه چیزی را اثبات نمی کند

- نشان نمی دهد AI پیامد بیمار را بهتر کرد. در endpoint اصلی ۱۴ روزه منفعت معنادار نبود.
- نشان نمی دهد AI بی فایده است. نقطه برآورد به نفع آن بود، مستندسازی در همه حوزه ها بهتر شد و هزینه دارو رو به پایین رفت؛ نتیجه null با منفعت کوچک واقعی که مطالعه برای تأییدش کوچک بوده سازگار است.
- ثابت نمی کند ابزار برای آسیب های نادر ایمن است. سیگنال ایمنی دیده نشد، اما مطالعه برای تأیید ایمنی رویدادهای شدید کمپاب طراحی نشده بود.
- نشان نمی دهد AI «از پزشک بهتر است» یا clinician را جایگزین می کند. این پشتیبان تصمیم است؛ اختیار کامل با clinician ماند.
- خودکار قابل تعمیم نیست. کارآزمایی در یک شبکه خصوصی شهری در کنیا بود؛ محیط روستایی، پیرامون شهری یا کشورهای پردرآمد ممکن است در هر جهت متفاوت باشند.
- صرفه جویی هزینه را تثبیت نمی کند. سیگنال هزینه پیشنهادکننده است، نه ارزیابی اقتصادی کامل.

قدرت شواهد چقدر است؟

برای ادعای مرکزی — نبود کاهش اثبات شده در آسیب کوتاه مدت بیمار — طراحی شواهد قوی است و نتیجه گیری به درستی فروتنانه. کارآزمایی prospective، از پیش ثبت شده، cluster-randomized با پیامد ترکیبی بیمارمحور و داوری کور، نزدیک به بهترین نوع شاهد واقعی است که برای چنین ابزاری می توان جمع کرد. بسیار اطلاعاتی تر از نمرة benchmark یا مطالعه vignette است.

برای یافته های ثانویه — مستندسازی بهتر، نسخه نویسی بدون تغییر، رضایت مشابه، هزینه آنتی بیوتیک پایین تر — شواهد خوب اند اما باید ثانویه خوانده شوند: سیگنال های حمایتی، نه تیترا، و همچنان حساس به آلودگی دو بازو و محدودیت تک شبکه ای.

برای ایمنی، شاهد آزمایش بخش اما محدود است: سیگنالی پیدا نشد، اما مطالعه برای یافتن آسیب نادر ساخته نشده بود.

مفیدترین موضع «نه کار می کند» است و نه «شکست خورد». این است: یک کارآزمایی واقعی و دقیق نشان داد این ابزار AI هیچ سیگنال ایمنی ایجاد نکرد و فرایند مراقبت را بهتر کرد، بدون این که منفعتی برای پیامد بیمار در دو هفته نشان دهد — و برای دیدن منفعت احتمالی به مطالعه ای بسیار بزرگ تر نیاز است.

چرا مهم است

بحث هوش مصنوعی پزشکی دقیقاً از چنین شواهدی کم دارد. هزاران مقاله وجود دارد که مدل ها در امتحان عالی اند یا در پرونده های مرتب با clinician برابری می کنند. کارآزمایی pragmatic بزرگ و تصادفی که پرسد بیمار واقعی بهتر شد بسیار کم است. این یکی از آن هاست و روی حقیقت کم زرق و برق فروود می آید: قبولی در آزمون با کمک کردن به بیمار یکی نیست.

همین شکاف کل داستان است. ابزار می تواند واقعاً برای clinician مفید باشد — یادداشت روشن تر، قبض داروی کمتر، یک جفت چشم دوم — و در عین حال ظرف دو هفته endpoint سخت بیمار را جابه جا نکند. هر دو می توانند هم زمان درست باشند و نظام سلامت بالغ باید هر دو را نگه دارد، نه فقط یکی را که به نفعش است.

این مقاله بار اثبات را هم مفید تنظیم می کند. اگر شرکتی می خواهد ادعا کند AI بالینی اش مراقبت را بهتر می کند، شاهد مرتبط leaderboard نیست؛ کارآزمایی ای مانند این است، روی پیامدهایی که برای بیمار مهم اند — و ترجیحاً بزرگ تر، چون درس صادقانه این جا این است که برای دیدن منفعت متوسط به مطالعه بزرگ نیاز دارید.

خلاصه تمیز

یک کارآزمایی pragmatic و cluster-randomized در ۱۶ مرکز مراقبت اولیه کنیا، ابزار پشتیبان تصمیم‌گیری AI مولد AI («AI Consult») را که به پرونده الکترونیک clinical officerها افزوده شده بود آزمود. در میان ۹۶۹۱ بیمار، ترکیب دآوری شده شکست درمان در ۱۴ روز، ۲۲٪ با ابزار در برابر ۲۰٪ بدون آن بود (adjusted OR ۷۷.۰، ۹۵٪ CI ۵۵.۰-۸۰.۱، $P = 0.013$) — تفاوت معناداری وجود نداشت. ابزار هیچ سیگنال ایمنی نشان نداد، مستندسازی بالینی را در همه حوزه‌های نمره‌گذاری شده بهتر کرد، نسخه‌نویسی را تغییر نداد، رضایت بیمار را ثابت گذاشت و با هزینه اندکی کمتر آنتی‌بیوتیک همراه بود. نویسندگان نتیجه می‌گیرند در این محدودیت‌ها ایمن بود اما شکست درمان را کاهش نداد و هر منفعتی احتمالاً متوسط است؛ تشخیص اثری به اندازه مشاهده‌شده به مطالعه‌ای بسیار بزرگ‌تر، در حد ۱۰۰٬۰۰۰ بیمار، نیاز دارد. نتیجه نشان نمی‌دهد AI بالینی پیامد بیمار را بهتر می‌کند و نشان هم نمی‌دهد بی‌فایده است — نشان می‌دهد ابزاری که ایمن به نظر رسید و فرایند را بهتر کرد، در یک شبکه خصوصی شهری ظرف دو هفته منفعت قابل اثباتی برای بیمار ایجاد نکرد.

بررسی بی‌پرده

مقاله چه نشان می‌دهد: در یک کارآزمایی تصادفی واقعی، افزودن ابزار پشتیبان تصمیم‌مبنتی بر LLM به پرونده مراقبت اولیه هیچ سیگنال ایمنی ایجاد نکرد، کیفیت مستندسازی را بهتر کرد، نسخه‌نویسی را تغییر نداد و ترکیب سخت‌گیرانه شکست درمان ۱۴ روزه را به‌طور معنادار کاهش نداد.

چه چیزی محتمل است اما اثبات نشده: ابزار کاهش کوچک واقعی در شکست درمان ایجاد می‌کند که این مطالعه برای دیدنش کوچک بوده؛ پس از حساب کردن همه هزینه‌ها پول صرفه‌جویی می‌کند؛ یا مستندسازی بهتر در بلندمدت به مراقبت بهتر تبدیل می‌شود.

چه چیزی را نشان نمی‌دهد: AI بالینی پیامد بیماران را بهتر می‌کند؛ برای رویدادهای نادر ناایمن است؛ clinician را جایگزین می‌کند یا از او بهتر است؛ نتیجه به محیط روستایی یا پردرآمد منتقل می‌شود؛ یا صرفه‌جویی هزینه قطعی است.

محدودیت‌های اصلی: مطالعه برای اثر بزرگ‌تری از اثر مشاهده‌شده توان‌دهی شده بود؛ یک خطای پیکربندی بازوی کنترل را آلوده کرد و عادت‌های شبکه مشترک نیز مقایسه را مخو می‌کنند، هر دو به نفع «تفاوت صفر»؛ یک شبکه خصوصی شهری با استاندارد نسبتاً بالا، نبود چارچوب از پیش تعیین‌شده noninferiority یا ایمنی؛ افق کوتاه ۱۴ روزه.

یک خواننده عمومی چقدر باید مطمئن باشد؟ اطمینان بالا که ابزار در این کارآزمایی سیگنال ایمنی نشان نداد و مستندسازی را بهتر کرد. اطمینان بالا که این‌جا منفعت قابل‌اثباتی برای پیامد ۱۴ روزه بیمار نشان نداد. اطمینان پایین به هر ادعای «کار می‌کند» یا «شکست خورده» به‌عنوان مداخله بیمارمحور — آن پرسش واقعاً باز است و مطالعه بسیار بزرگ‌تری می‌خواهد. موضع مناسب: نتیجه‌ای دقیق و واقعی درباره ابزاری که ایمن به نظر رسید و فرایند مراقبت را بهتر کرد، نه حکم نهایی درباره این که AI مراقبت اولیه را متحول یا خراب می‌کند.

منابع

- (2026) Medicine Nature al., et Agweyu
<https://doi.org/10.1038/s41591-026-04503-6>
- 202502499779176 registration Registry, Trials Clinical Pan-African
<https://pactr.samrc.ac.za/>
- (2026) trial the on release press PATH / EurekaAlert
<https://www.eurekaalert.org/news-releases/1133583>