The Clean Paper

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

Eine klinische KI, die sicher wirkte und die Dokumentation verbesserte — aber nicht die Patientenergebnisse

Eine pragmatische, cluster-randomisierte Studie in 16 kenianischen Einrichtungen der Primärversorgung testete ein generatives KI-Entscheidungsunterstützungssystem in der elektronischen Patientenakte der Clinical Officers. Unter 9.691 Patienten lag der strenge, von Experten adjudizierte kombinierte 14-Tage-Endpunkt für Behandlungsversagen bei 2,2 % mit dem Tool gegenüber 2,0 % ohne (adjustierte Odds Ratio 0,77; 95%-KI 0,55–1,08; P = 0,13) — kein signifikanter Unterschied. Das Tool zeigte kein Sicherheitssignal, verbesserte die Dokumentation in allen bewerteten Bereichen, ließ Verschreibung und Zufriedenheit unverändert und tendierte zu niedrigeren Antibiotikakosten. Das ist keine gescheiterte Studie, sondern eine seltene ehrliche — an Patienten statt an Benchmarks gemessen — und sie landet bei der unspektakulären Wahrheit, dass ein Tool, das sicher wirkte und den Prozess verbesserte, Patienten innerhalb von zwei Wochen nicht nachweisbar half; ein möglicher Nutzen dieser Größenordnung würde eine viel größere Studie erfordern.

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

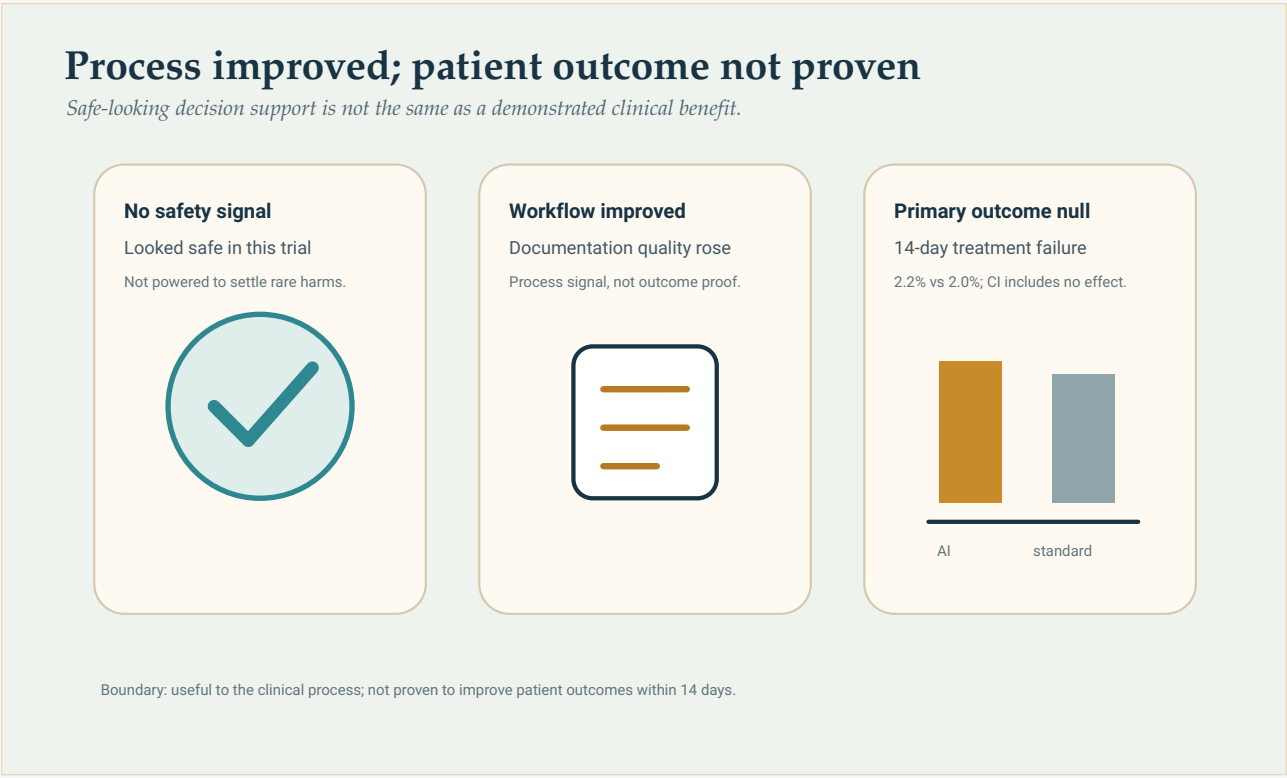
Sicher und hilfreich ist nicht dasselbe wie wirksam

Die meisten Schlagzeilen über medizinische KI entstehen aus der falschen Art von Studie. Ein Modell schneidet bei Prüfungsfragen gut ab oder schlägt Ärzte bei sorgfältig ausgewählten Fallvignetten – und die Geschichte schreibt sich von selbst: Die Maschine ist bereit für die Klinik.

Diese Studie tat etwas Schwierigeres und Selteneres. Sie brachte ein generatives KI-Entscheidungsunterstützungs in echte Primärversorgung, in reale Einrichtungen, zu echten Patienten, und stellte die Frage, auf die es wirklich ankommt: Ging es den Patienten besser?

Die ehrliche Antwort lautet nein – jedenfalls nicht messbar innerhalb von 14 Tagen. Das Tool zeigte kein Sicherheitssignal. Es verbesserte die Qualität der klinischen Dokumentation. Möglicherweise senkte es sogar einige Medikamentenkosten. Aber es verringerte Behandlungsversagen nicht signifikant, und die Autoren sagen vorsichtig, dass ein möglicher Nutzen wahrscheinlich nur moderat ist.

Das ist kein Scheitern der Studie. Das ist die Studie, die ihren Zweck erfüllt. So sieht verantwortungsvolle Evidenz über KI in der Medizin aus, wenn sie an Patienten statt an Benchmarks gemessen wird.



Das Tool zeigte kein Sicherheitssignal und verbesserte die klinische Dokumentation, aber der vorab festgelegte 14-Tage-Patientenendpunkt verbesserte sich nicht signifikant. Prozesshilfe ist nicht dasselbe wie nachgewiesener Patientennutzen.

Was die Autoren gemacht haben

Das Team führte eine pragmatische, cluster-randomisierte Studie in **16 Einrichtungen der Primärversorgung** eines privaten Gesundheitsnetzes (Penda Health) in den kenianischen Counties Nairobi und Kiambu durch. Die Versorgung wird dort größtenteils von **Clinical Officers** geleistet – medizinischen Fachkräften mittlerer Qualifikationsstufe mit dreijährigem Diplom –, häufig ohne einfachen Zugang zu einer ärztlichen Senior-Konsultation.

Randomisiert wurde auf Ebene der Behandler, nicht der Patienten. **103 Clinical Officers** wurden zugeteilt: 52 zur Interventions- und 51 zur Kontrollgruppe. Beide Gruppen verwendeten dieselbe cloudbasierte elektronische Patientenakte. In der Interventionsgruppe kam zusätzlich **„AI Consult"** (Version 2.0) hinzu, ein auf dem großen Sprachmodell **GPT-4o von OpenAI** basierendes Entscheidungsunterstützungssystem, das direkt in die Akte eingebettet war. Es las die vom Behandler dokumentierten Informationen und konnte mögliche Probleme bei Diagnose oder Therapieplan markieren. Die Behandler behielten volle Autonomie: Sie konnten Vorschläge annehmen, verändern oder ignorieren.

> **Welches Modell war es, und warum die Details wichtig sind**
>
> Das Tool war **AI Consult 2.0** und lief mit **GPT-4o von OpenAI** (Release vom Mai 2025), angesprochen über OpenAIs kommerzielle API unter einer Enterprise-Lizenz und mit geringer Zufälligkeit (Temperatur 0,1). Es war in eine eigens angepasste elektronische Patientenakte (EasyClinic EMR) eingebettet und wurde durch Systemprompts auf die nationalen kenianischen Behandlungsleitlinien ausgerichtet; die Autoren veröffentlichten den vollständigen Instruktionsprompt.
> Warum diese Details? Weil das Ergebnis *ein bestimmtes System* betrifft – eine Modellversion, einen Prompt, eine Patientenakte, eine Konfiguration – und nicht „LLMs in der Medizin" allgemein. Die Autoren machen denselben Punkt: Sie nennen ihr Ergebnis einen *zeitgebundenen Benchmark und keine feste Schätzung der Leistungsfähigkeit*. Ein neueres Modell, ein anderer Prompt oder eine weniger digitalisierte Klinik könnten das Ergebnis verändern.
> Zur Unabhängigkeit: OpenAI stellte später Sachleistungen bereit (Cloud-Computing-Credits und technische Hinweise zur Nutzung der API). Die Autoren geben jedoch an, dass die Entscheidung für OpenAI *vor* diesem Angebot getroffen wurde und dass OpenAI weder am Studiendesign noch an Datenerhebung, Analyse oder Publikationsentscheidung beteiligt war.

Zwischen dem 22. April und dem 16. Juli 2025 wurden **9.691 Patienten** eingeschlossen. Der **primäre Endpunkt war bewusst patientenzentriert und streng**: ein von Experten adjudizierter *kombinierter Endpunkt für Behandlungsversagen* innerhalb von 14 Tagen nach dem Besuch. Ein Gremium von Ärzten bewertete verblindet gegenüber der Studiengruppe, ob ein Patient ein ungünstiges Ergebnis wie eine nicht abgeklungene oder sich verschlechternde Erkrankung hatte. Die Studie war im Voraus registriert (Pan-African Clinical Trials Registry 202502499779176).

Genau diese Designentscheidung ist der Punkt. Es ist leicht zu zeigen, dass ein KI-Tool verändert, was ein Behandler dokumentiert. Viel schwieriger – und viel bedeutsamer – ist zu zeigen, dass es verändert, was mit dem Patienten passiert.

Was sie herausfanden

Der primäre Endpunkt verbesserte sich nicht. Behandlungsversagen trat bei 102 von 4.693 Patienten (2,2 %) in der KI-Gruppe und bei 94 von 4.654 Patienten (2,0 %) in der Kontrollgruppe auf. Die rohen Prozentwerte waren mit KI minimal höher, doch nach statistischer Anpassung zeigte die Punktschätzung in Richtung eines Nutzens: Die **adjustierte Odds Ratio betrug 0,77 (95%-Konfidenzintervall 0,55 bis 1,08; P = 0,13)** – nicht statistisch signifikant. Dass Roh- und adjustierte Werte in unterschiedliche Richtungen weisen, ist kein Rechenfehler; die Anpassung berücksichtigt Unterschiede zwischen den Behandler-Clustern. So oder so umfasst das Konfidenzintervall deutlich „keinen Effekt", sodass kein Nutzen behauptet werden kann; absolut war der Unterschied sehr klein.

Eine allgemeinverständliche Erklärung dazu, wie Odds Ratios, Konfidenzintervalle und P-Werte zusammen gelesen werden, steht in unserem Leitfaden zu klinischen Ergebnissen.

Kein Sicherheitssignal – innerhalb klarer Grenzen. Kein schwerwiegendes unerwünschtes Ereignis wurde dem Tool zugeschrieben, und eine unabhängige Begutachtung fand kein Sicherheitssignal. Die Autoren benennen die Grenze dieser Beruhigung offen: Die Studie war nicht dafür gepowert, seltene schwere Schäden zu erkennen, und hatte weder eine vorab festgelegte Nichtunterlegenheitsanalyse noch ein formales Sicherheitsframework. Sie kann Sicherheit bei seltenen Ereignissen daher nicht *beweisen*.

Die Dokumentation wurde besser. Bei 2.000 Begegnungen, die verblindete Experten überprüften, dokumentierten Behandler mit AI Consult in allen bewerteten Bereichen besser – bei der festgehaltenen Diagnose, beim Behandlungsplan und bei der Vollständigkeit insgesamt.

Die Verschreibungspraxis änderte sich kaum. Es gab keinen signifikanten Unterschied beim Verschreiben, einschließlich korrekter Antibiotikaverwendung (adjustierte Odds Ratio 0,86; 95%-KI 0,48 bis 1,55). Die Rate von Antibiotikaverschreibungen veränderte das Tool nicht.

Die Patienten bemerkten keinen Unterschied. Unter 826 Patienten, die eine Zufriedenheitsbefragung abschlossen, war die Zufriedenheit in beiden Gruppen praktisch identisch; auch die Konsultationsdauer war ähnlich.

Die Kosten deuteten leicht nach unten. In einer adjustierten Analyse waren antibiotikabezogene Kosten in der KI-Gruppe niedriger – vermutlich durch günstigere Auswahl statt weniger Verschreibungen – und die Einsparung pro Patient schien größer als die Kosten des Toolbetriebs pro Patient. Die Autoren kennzeichnen dies als Hinweis, nicht als abschließendes Ergebnis: Eine vollständige Total-Cost-of-Ownership-Analyse lag außerhalb des Studienumfangs.

Die Ein-Satz-Zusammenfassung der Autoren ist am saubersten: LLM-Unterstützung war innerhalb dieser Grenzen sicher, verringerte aber Behandlungsversagen in 14 Tagen nicht; ein möglicher Nutzen dürfte gering sein.

Warum „kein signifikanter Unterschied“ nicht „es funktioniert nicht“ bedeutet

Ein nuller primärer Endpunkt lässt sich leicht in beide Richtungen überinterpretieren. Zwei Punkte verhindern die einfache Geschichte.

Erstens war **die Studie dafür geplant, einen größeren Effekt zu erkennen, als sie beobachtete**. Schwerwiegende schlechte Ergebnisse sind in der Primärversorgung selten – hier etwa 2 %. Einen kleinen realen Nutzen vom Rauschen zu trennen, braucht deshalb enorme Fallzahlen. Die nachträglichen Power-Berechnungen der Autoren deuten darauf hin, dass ein Effekt in der beobachteten Größenordnung eine **wesentlich größere Studie mit mehr als 100.000 Patienten** erfordern würde. Ein nicht signifikantes Ergebnis in einer Studie dieser Größe schließt einen kleinen realen Nutzen nicht aus; es bedeutet, dass diese Studie ihn nicht auflösen konnte.

Zweitens war **der Vergleich teilweise verwischt**. Ein Konfigurationsfehler verschaffte einigen Behandlern der Kontrollgruppe kurzzeitig Zugang zu AI Consult, und Mitarbeitende im selben Netzwerk sprechen miteinander und tragen Gewohnheiten über Gruppengrenzen hinweg. Beides macht die beiden Arme ähnlicher und drückt einen möglichen echten Unterschied in Richtung null. Zudem arbeitete das Netzwerk bereits auf relativ hohem Niveau, sodass weniger Raum für sichtbare Verbesserung bleibt.

Nichts davon rettet eine „Durchbruch“-Schlagzeile. Aber es bedeutet, dass die richtige Lesart kalibriert und nicht bloß abwertend ist: Beim härtesten und ehrlichsten Endpunkt half dieses Tool Patienten innerhalb von zwei Wochen nicht nachweisbar – während es die Dokumentation messbar verbesserte und kein Sicherheitssignal zeigte.

Was dies *nicht* beweist

- Es zeigt **nicht**, dass die KI Patientenergebnisse verbessert hat. Beim primären 14-Tage-Endpunkt gab es keinen signifikanten Nutzen.
- Es zeigt **nicht**, dass die KI nutzlos ist. Die Punktschätzung begünstigte sie, die Dokumentation verbesserte sich durchgehend und Medikamentenkosten tendierten nach unten; das Nullergebnis ist mit einem kleinen realen Nutzen vereinbar, für dessen Nachweis die Studie zu klein war.
- Es beweist **nicht**, dass das Tool bei seltenen Schäden sicher ist. Es gab kein Sicherheitssignal, doch die Studie war weder groß genug noch dafür konzipiert, seltene schwere Ereignisse zuverlässig auszuschließen.
- Es zeigt **nicht**, dass „KI Ärzte schlägt“ oder Behandler ersetzt. Es handelt sich um Entscheidungs*unterstützung*; die Behandler behielten die volle Entscheidungshoheit.
- Es lässt sich **nicht automatisch verallgemeinern**. Die Studie lief in einem einzigen privaten urbanen Netzwerk in Kenia; ländliche, periurbane oder einkommensstärkere Settings könnten in beide Richtungen abweichen.
- Es belegt **keine Kosteneinsparung**. Das Kostensignal ist suggestiv, keine vollständige gesundheitsökonomische Evaluation.

Wie belastbar sind die Belege?

Für die zentrale Aussage – keine nachgewiesene Verringerung kurzfristiger patientenbezogener Schäden – ist die Evidenz *vom Design her stark* und *in der Schlussfolgerung angemessen bescheiden*. Eine prospektive, vorregistrierte, cluster-randomisierte Studie mit verblindetem patientenbezogenem kombinierten Endpunkt gehört zu den besten Real-World-Evidenzen, die sich für ein solches Tool sammeln lassen. Sie ist wesentlich informativer als ein Benchmark-Score oder eine Vignettenstudie.

Für die sekundären Befunde – bessere Dokumentation, unveränderte Verschreibung, ähnliche Zufriedenheit, niedrigere Antibiotikakosten – ist die Evidenz gut, sollte aber als sekundär gelesen werden: unterstützende Signale, nicht die Hauptaussage, und anfällig für dieselbe Kontamination und die Begrenzung auf ein Netzwerk.

Bei Sicherheit ist die Evidenz beruhigend, aber begrenzt: kein Signal gefunden, aber keine Studie, die seltene Schäden zuverlässig finden sollte.

Die nützlichste Haltung lautet deshalb weder „es funktioniert“ noch „es ist gescheitert“. Sie lautet: Eine sorgfältige Real-World-Studie fand bei diesem KI-Tool kein Sicherheitssignal und eine Verbesserung des Versorgungsprozesses, ohne innerhalb von zwei Wochen einen Patientennutzen nachzuweisen – und ein solcher Nutzen würde, falls klein, eine sehr viel größere Studie erfordern.

Warum das wichtig ist

Der Debatte über medizinische KI fehlt genau diese Art von Evidenz. Es gibt Tausende Arbeiten, in denen Modelle Prüfungen bestehen oder bei sauberen Fallvignetten mit Ärzten mithalten. Es gibt nur sehr wenige große, pragmatische, randomisierte Studien, die messen, ob es realen Patienten besser geht. Diese ist eine davon, und sie landet bei der unspektakulären Wahrheit: Einen Test zu bestehen ist nicht dasselbe wie einem Patienten zu helfen.

Diese Lücke ist die ganze Geschichte. Ein Tool kann für Behandler wirklich nützlich sein – klarere Notizen, niedrigere Medikamentenkosten, ein zweites Augenpaar – und trotzdem einen harten Patientenausgang innerhalb von zwei Wochen nicht verändern. Beides kann gleichzeitig wahr sein, und ein reifes Gesundheitssystem muss beide Tatsachen zusammenhalten, statt sich die bequemere herauszupicken.

Die Arbeit setzt außerdem die Beweislast in eine nützliche Richtung zurück. Wenn ein Unternehmen behaupten will, seine klinische KI verbessere Versorgung, ist die relevante Evidenz keine Rangliste. Es ist eine Studie wie diese mit Endpunkten, die Patienten etwas bedeuten – und idealerweise eine größere, denn die ehrliche Lektion hier lautet: Moderate Vorteile brauchen große Studien, damit man sie sehen kann.

Kurz zusammengefasst

Eine pragmatische, cluster-randomisierte Studie in 16 kenianischen Einrichtungen der Primärversorgung testete ein generatives KI-Entscheidungsunterstützungssystem („AI Consult“), das der elektronischen Patientenakte der Clinical Officers hinzugefügt wurde. Unter 9.691 Patienten lag der von Experten adjudizierte kombinierte Endpunkt für Behandlungsversagen innerhalb von 14 Tagen bei 2,2 % mit dem Tool gegenüber 2,0 % ohne (adjustierte Odds Ratio 0,77; 95%-KI 0,55–1,08; P = 0,13) – kein signifikanter Unterschied. Das Tool zeigte kein Sicherheitssignal, verbesserte die klinische Dokumentation in allen bewerteten Bereichen, veränderte die Verschreibungspraxis nicht, ließ die Patientenzufriedenheit unverändert und war mit etwas niedrigeren Antibiotikakosten verbunden. Die Autoren schließen, dass es innerhalb dieser Grenzen sicher war, aber Behandlungsversagen nicht verringerte und ein möglicher Nutzen vermutlich moderat ist; um einen Effekt der beobachteten Größe zu erkennen, wäre eine sehr viel größere Studie in der Größenordnung von 100.000 Patienten nötig. Das Ergebnis zeigt weder, dass klinische KI Patientenergebnisse verbessert, noch dass sie nutzlos ist – es zeigt, dass ein Tool, das sicher wirkte und den Prozess verbesserte, Patienten innerhalb von zwei Wochen in einem privaten urbanen Netzwerk nicht nachweisbar half.

Nüchtern geprüft

Was das Paper zeigt: In einer randomisierten Real-World-Studie erzeugte ein LLM-basiertes Entscheidungsunterstützungssystem in der Primärversorgung kein Sicherheitssignal, verbesserte die Qualität klinischer Dokumentation, veränderte die Verschreibung nicht und reduzierte einen strengen 14-Tage-Patientenendpunkt für Behandlungsversagen nicht signifikant.

Was plausibel, aber nicht bewiesen ist: Dass das Tool eine kleine reale Verringerung des Behandlungsversagens bewirkt, die diese Studie nicht erkennen konnte; dass es nach vollständiger Kostenrechnung Geld spart; dass bessere Dokumentation langfristig zu besserer Versorgung führt.

Was es nicht zeigt: Dass klinische KI Patientenergebnisse verbessert; dass sie für seltene Ereignisse sicher ist; dass sie Behandler ersetzt oder übertrifft; dass diese Ergebnisse auf ländliche oder einkommensstärkere Settings übertragbar sind; dass Kosteneinsparungen gesichert sind.

Wichtigste Einschränkungen: Gepowert für einen größeren Effekt als beobachtet (seltene Endpunkte brauchen sehr große Stichproben); ein Konfigurationsfehler kontaminierte die Kontrollgruppe, und Gewohnheitstransfer im gemeinsamen Netzwerk verwischt den Vergleich – beides verzerrt in Richtung null; ein einziges privates urbanes Netzwerk mit bereits hohem Standard; kein vorab festgelegtes Nichtunterlegenheits- oder Sicherheitsframework; kurze Nachbeobachtung von 14 Tagen.

Wie viel Vertrauen sollte ein allgemeines Publikum haben? Hoch, dass das Tool in dieser Studie kein Sicherheitssignal zeigte und die Dokumentation verbesserte. Hoch, dass es hier keinen nachweisbaren 14-Tage-Patientennutzen gab. Niedrig für die pauschale Aussage, es „funktioniere“ oder „funktioniere nicht“ als Intervention für Patientennutzen – genau diese Frage bleibt offen und braucht eine viel größere Studie. Angemessene Haltung: ein sorgfältiges Real-World-Ergebnis über ein Tool,

das kein Sicherheitssignal zeigte und den Versorgungsprozess verbesserte, kein Urteil, dass KI die Primärversorgung revolutioniert – oder ruiniert.

Quellen

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