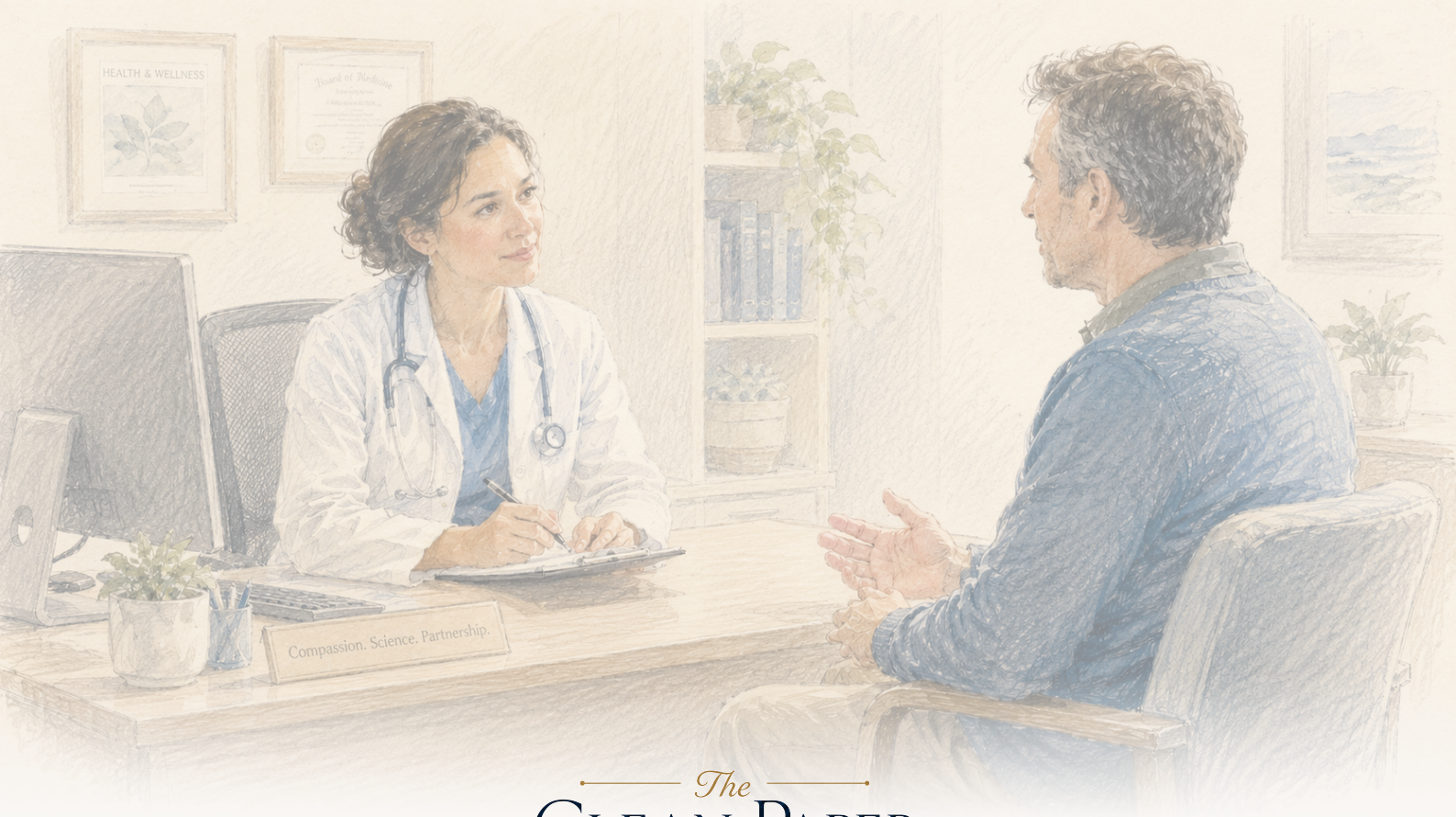


The Clean Paper

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

安全そうで事務作業は改善した臨床 AI—— しかし患者の転帰は改善しなかった

ケニアの 16 のプライマリケア施設で、*clinical officers* が使用する記録に生成 *AI* による意思決定支援ツールを追加し、実用的なクラスター無作為化試験で検証した。*9,691* 人の患者において、専門家が判定した厳格な *14* 日間の治療失敗の複合評価項目は、ツールありで *2.2%*、なしで *2.0%* だった（調整オッズ比 *0.77*、*95% CI 0.55-1.08*、*P = 0.13*）。有意差はなかった。ツールは安全性シグナルを示さず、評価されたすべての領域で記録を改善し、処方と満足度は変わらず、抗菌薬費用は低下する方向を示した。これは失敗した研究ではない。患者ではなくベンチマークで評価するのではなく、患者を対象に測定した、稀で正直な研究である。そして、安全そうに見え、プロセスを助けたツールが *2* 週間で患者を明らかに助けたとはいえず、そのような利益を検出するにははるかに大規模な試験が必要だという、華やかではない真実に行き着く。

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

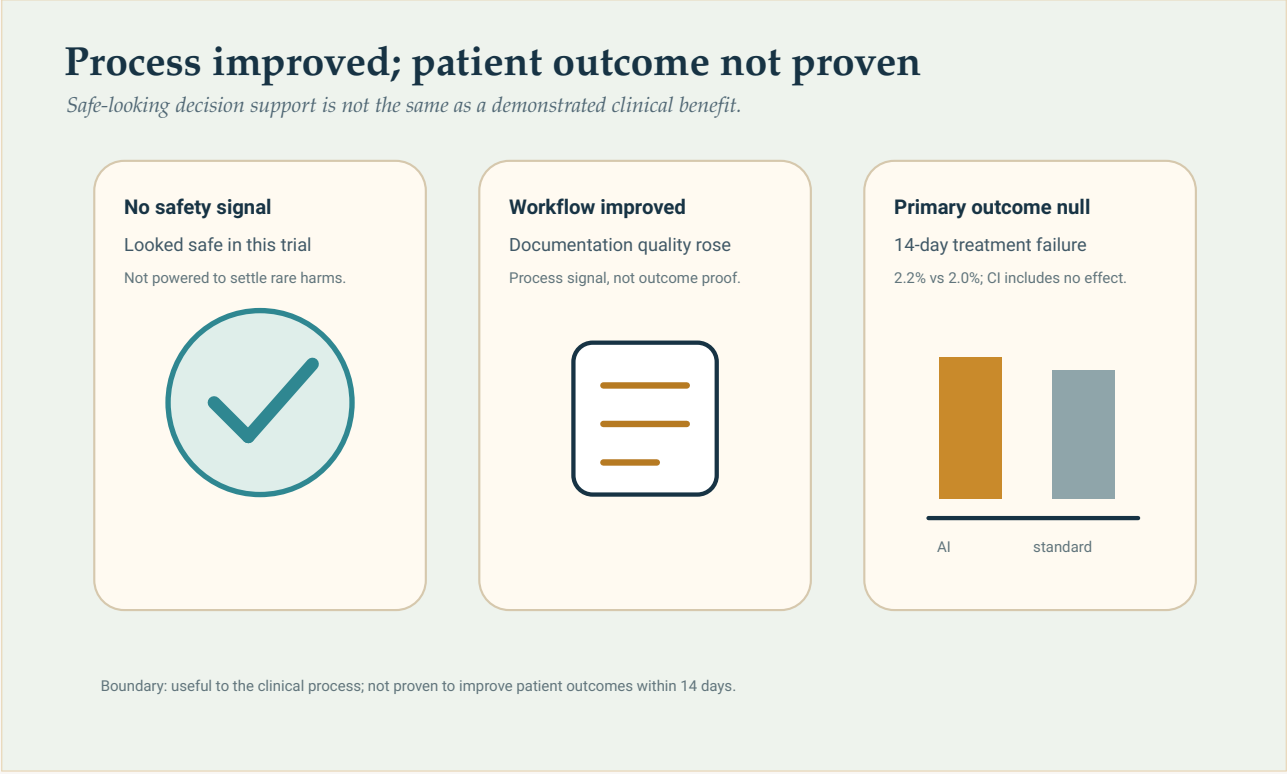
安全で役立つことは、有効であることと同じではない

医療 AI をめぐる見出しの多くは、研究としては筋違いの種類のものから生まれている。モデルが試験問題で高得点を取ったり、選別された症例提示で医師を上回ったりすると、話は決まったように展開する。機械は臨床現場に出る準備ができた、と。

この試験が試みたのは、より難しく、より稀なことだった。生成 AI による意思決定支援ツールを、実際のプライマリケア、実際の医療施設、実際の患者に導入し、本当に重要な問いを投げかけた。患者の転帰は改善したのか、と。

正直な答えは、そうではない —— 少なくとも、14 日間では測定可能な形で改善しなかった。ツールは安全性シグナルを示さなかった。臨床記録の質は改善した。一部の薬剤費さえ下がった可能性がある。しかし、治療失敗を有意に減らすことはなく、著者らは、利益が存在するとしてもおそらく小さいと慎重に述べている。

これは研究の失敗ではない。研究が機能したということだ。ベンチマークではなく患者を対象に測定したとき、医療 AI に関する責任あるエビデンスとはこのようなものになる。



このツールは安全性シグナルを示さず、臨床記録を改善したが、事前に規定された 14 日間の患者転帰は有意に改善しなかった。プロセスの支援は、患者に対する利益が証明されたことと同じではない。

The Clean Paper CC BY 4.0

著者らが行ったこと

チームは、ケニアのナイロビ郡とキアンブ郡にある、民間医療ネットワーク（Penda Health）が運営する**16 のプライマリケア施設**で、実用的なクラスター無作為化試験を実施した。これらの施設での診療は主に**クリニカル・オフィサー（clinical officers）**—— 三年間のディプロマを取得した中級医療従事者 —— が担っており、上級者に気軽に相談できないことも多い。

無作為化の単位は患者ではなく臨床医だった。**103 人のクリニカル・オフィサー**が無作為化され、介入群 52 人、対照群 51 人となった。両群とも同じクラウドベースの電子カルテを使用した。介入群にはさらに、その記録に組み込まれた意思決定支援ツール「**AI Consult**」（version 2.0）が提供された。このツールは **OpenAI の GPT-4o** 大規模言語モデルを基盤としていた。臨床医が記録した情報を読み取り、診断や治療計画に関する潜在的な問題を指摘できた。臨床医の自律性は完全に保たれ、提案を受け入れることも、変更することも、無視することもできた。

> **どのモデルだったのか、そして具体的な情報が重要な理由**
>
> このツールは **AI Consult 2.0** で、**OpenAI の GPT-4o**（2025 年五月リリース）を稼働させ、エンタープライズライセンスの下で OpenAI の商用 API を通じて利用していた。ランダム性を低くした設定（temperature 0.1）で運用されていた。EasyClinic の EMR という特注の電子カルテ内に組み込まれ、ケニアの国の治療ガイドラインに沿うよう作成されたシステムプロンプトによって指示されていた。著者らは、その完全な指示プロンプトを公開している。
> なぜここまで明記するのか。結果が示しているのは、「**医療における LLM**」全般ではなく、特定の一つのシステム —— 一つのモデルバージョン、一つのプロンプト、一つの記録システム、一つの環境 —— についてだからだ。著者らも同じ点を指摘しており、自分たちの発見を、能力の固定的な推定値ではなく、その時点でのベンチマークと呼んでいる。より新しいモデル、異なるプロンプト、あるいはデジタル化が進んでいない診療所なら、結果はいずれの方向にも動きうる。
> 独立性については、OpenAI は後に現物支援（クラウド計算資源のクレジットと、API 利用に関する技術的助言）を提供した。しかし著者らによれば、OpenAI を利用する決定はその申し出の前に行われており、OpenAI は試験の設計、データ収集、解析、または出版の決定に関与していない。

2025 年四月 22 日から七月 16 日までに、**9,691 人の患者**が登録された。**主要評価項目は、意図的に患者中心で厳格なものだった**。来院から 14 日以内の、専門家が判定する治療失敗の複合評価項目である。臨床医のパネルが研究群を知らされない状態で、未解決または悪化した疾患など、各患者に不良転帰があったかを判定した。試験は事前に登録されていた（Pan-African Clinical Trials Registry 202502499779176）。

この設計上の選択こそがポイントだ。AI ツールによって臨床医の記録内容が変わったことを示すのは容易だ。患者に実際に起きることを変えたと示すのは、はるかに難しく、またはるかに意味がある。

何が分かったのか

主要評価項目は改善しなかった。治療失敗は、AI 群では 4,693 人中 102 人（2.2%）、対照群では 4,654 人中 94 人（2.0%）に発生した。未調整の割合は AI 群でわずかに高かったが、調整後の点推定値は利益の方向を示した。**調整オッズ比は 0.77（95% 信頼区間 0.55〜1.08、P = 0.13）**で、統計的に有意ではなかった。未調整値と調整後の値がこのように逆転して見えるのは計算ミスではなく、調整によって臨床医クラスター間の差を考慮しているためだ。いずれにせよ、信頼区間には「効果なし」が十分に含まれているため、利益を主張することはできず、絶対値で見ても効果はごく小さい。

オッズ比、信頼区間、P 値を合わせてどう読むかについては、臨床結果の読み方ガイドを参照されたい。

安全性シグナルはなかった —— ただし限界がある。ツールに関連すると判定された重篤な有害事象はなく、独立したレビューでも安全性シグナルは見つからなかった。著者らは、この安心材料には限界があることも正直に述べている。この試験は、まれな重篤な有害事象を検出できるよう設計されておらず、事前に規定された非劣性の枠組みや正式な安全性評価の枠組みもなかった。そのため、まれな事象について安全性を証明することはできない。

記録は改善した。盲検化された専門家が評価した 2,000 件の診療のうち、AI Consult を使った臨床医は、評価対象となったすべての領域で、より良い臨床記録を作成した。記録された診断、治療計画、全体的な完全性のいずれも改善した。

処方はほとんど変わらなかった。正しい抗菌薬使用を含む処方に、有意差はなかった（調整オッズ比 0.86、95% CI 0.48〜1.55）。このツールは抗菌薬の処方率を変えなかった。

患者は違いを感じなかった。満足度調査を完了した 826 人の患者では、満足度は両群で実質的に同じで、診療時間も同程度だった。

費用はわずかに低下する方向を示した。調整後の解析では、抗菌薬関連費用は AI 群で低かった。これは処方数が少なかったというより、より安価な選択による可能性がある。また、患者一人当たりの抗菌薬費用の節約額は、ツールの運用にかかる患者一人当たりの費用を上回るように見えた。著者らはこれを示唆的な結果であり、確定したものではないと注意している。総保有コストを完全に算定することは、この試験の範囲外だった。

著者ら自身による一文の要約が最も簡潔だ。LLM による支援はこれらの限界の範囲内では安全だったが、14 日以内の治療失敗を減らすことはなく、利益があるとしてもおそらく小さい。

「有意差なし」は「効かない」ではない理由

主要評価項目が無効だったという結果は、どちらの方向にも簡単に読み過ぎてしまう。単純な物語を止める要素が二つある。

第一に、**この試験は、今回見つかったものより大きな効果を捉えるよう設計されていた。**プライマリケアにおける深刻な不良転帰はまれで、この試験では約 2% だった。そのため、小さな真の利益とノイズを区別するには膨大な人数が必要になる。著者ら自身による事後的な検出力計算では、今

回観察された規模の効果を検出するには、**はるかに大規模な試験、すなわち 100,000 人を超える規模**が必要になることが示唆されている。この規模の研究で有意でなかった結果は、小さくても現実に存在する利益を否定するものではない。そのような利益を、この研究では解明できなかったということだ。

第二に、**比較が部分的にぼやけていた**。設定エラーによって、一部の対照群の臨床医が一時的に AI Consult を利用できる状態になった。また、同じネットワーク内の臨床医は互いに話し、境界を越えて習慣を持ち込む。いずれの影響も、二群をより似たものにする方向に働き、現実に存在する差をゼロへ近づける。そのうえ、実施施設のネットワークはもともと比較的高い水準で運営されており、ツールが改善を示す余地が小さかった。

これらの点をもって、「画期的進歩」という見出しを救うことはできない。しかし、正しい読み方は、意気消沈するものではなく、適切に調整されたものになるべきだ。最も難しく、最も正直な評価項目において、このツールは二週間で患者を明らかに助けたとはいえなかった。一方で、記録作成を測定可能な形で助け、安全そうに見えた。

これは何を証明していないのか

- AI が患者の転帰を改善したことを**示すものではない**。主要な 14 日間の評価項目では、有意な利益はなかった。
- AI が役に立たないことを**示すものではない**。点推定値は AI を支持し、記録は全般的に改善し、薬剤費は低下する方向を示した。無効という結果は、試験が確認するには小さすぎた、現実には小さな利益がある可能性と矛盾しない。
- まれな有害事象に対してツールが安全であることを**証明するものではない**。安全性シグナルは示されなかったが、まれな重篤事象について安全性を認証するような検出力も設計もなかった。
- 「AI が医師に勝つ」ことや、臨床医を置き換えることを**示すものではない**。これは意思決定の支援であり、臨床医はそれを受け入れるか拒否するかについて完全な権限を持っていた。
- 結果が自動的に一般化できることを**示すものではない**。試験はケニアの都市部にある単一の民間ネットワークで実施された。農村部、都市周辺部、高所得地域では、いずれの方向にも異なる結果になりうる。
- コスト削減を**確立するものではない**。費用に関するシグナルは示唆的なものであり、完全な経済評価ではない。

エビデンスの強さはどの程度か

中心的な主張 —— 短期的な患者の害を減らすことが証明されなかったという点 —— については、研究デザインとしては強く、結論としては適切に謙虚なエビデンスである。前向きに事前登録され、クラスター無作為化され、盲検化された患者レベルの複合評価項目を用いた試験は、このようなツールについて集められる現実世界のエビデンスとして、最良のものに近い。ベンチマークのスコアや症例提示研究より、はるかに多くの情報をもたらす。

二次的な知見 —— 記録の改善、処方の不変、同程度の満足度、抗菌薬費用の低下 —— については、

エビデンスは良好だが、二次的なものとして読むべきだ。中心的な結論ではなく、支持的なシグナルであり、同じ汚染や単一ネットワークという限界の影響を受けうる。

安全性については、エビデンスは安心材料になるが、範囲には限界がある。シグナルは見つからなかったが、まれな有害事象を見つけるよう設計された研究ではなかった。

最も有用な立場は、「効く」でも「失敗した」でもない。慎重な現実世界の試験によれば、この AI ツールは安全性シグナルを示さず、ケアのプロセスを改善したが、二週間で患者転帰の利益を示すことはなかった。そして、そのような利益を検出するには、はるかに大規模な研究が必要になる、という立場である。

なぜ重要なのか

医療 AI をめぐる議論には、まさにこの種のエビデンスが不足している。モデルが試験で高得点を取り、整然とした症例で臨床医に匹敵することを示す論文は何千本もある。しかし、実際の患者の状態が良くなったかを測定する、大規模で実用的な無作為化試験は非常に少ない。これはその一つであり、華やかではない真実に行き着く。試験に合格することは、患者を助けることと同じではない。

この隔たりこそが、話のすべてだ。ツールは臨床医にとって本当に役立つことがある。より明確な記録、より低い薬剤費、もう一組の目。それでも、二週間という期間では、厳格な患者転帰を動かさないことがある。両方の事実は同時に成り立ちうる。成熟した医療システムは、都合の良い一方を選ぶのではなく、両方を併せ持たなければならない。

また、これは立証責任を有用な方向に戻す。企業が自社の臨床 AI はケアを改善すると主張したいなら、関係するエビデンスはリーダーボードではない。患者にとって重要な転帰を対象にした、このような試験である。できれば、より大規模な試験が望ましい。今回の正直な教訓は、控えめな利益を見つけるには大規模な研究が必要だということだからだ。

簡潔な要約

ケニアの 16 のプライマリケア施設で、クリニカル・オフィサーが使用する電子カルテに生成 AI による意思決定支援ツール（「AI Consult」）を追加し、実用的なクラスター無作為化試験で検証した。9,691 人の患者において、専門家が判定した 14 日以内の治療失敗の複合評価項目は、ツールありで 2.2%、なしで 2.0% だった（調整オッズ比 0.77、95% CI 0.55–1.08、P = 0.13）。有意差はなかった。ツールは安全性シグナルを示さず、評価されたすべての領域で臨床記録を改善し、処方を変えず、患者満足度も変えず、抗菌薬費用はやや低かった。著者らは、これらの限界の範囲内では安全だったが、治療失敗を減らすことはなく、利益があるとしてもおそらく小さいと結論づけている。観察された規模の効果を検出するには、はるかに大規模な試験（100,000 人規模）が必要になる。この結果は、臨床 AI が患者転帰を改善することも、役に立たないことも示していない。安全そうに見え、プロセスを助けたツールが、都市部の単一の民間ネットワークにおいて、二週間で患者を明らかに助けたとはいえなかったことを示している。

冷静に見ると

論文が示していること：実際の医療現場での無作為化試験において、プライマリケアの記録に LLM ベースの意思決定支援ツールを追加しても、安全性シグナルは示されず、臨床記録の質は改善し、処方は変わらず、厳格な 14 日間の患者に関する治療失敗の複合評価項目も有意には減少しなかった。

もっともらしいが証明されていないこと：このツールが、今回の試験では検出できないほど小さい、治療失敗の現実的な減少をもたらすこと。全費用を算入すれば費用を節約すること。より良い記録が最終的により良いケアにつながること。

示していないこと：臨床 AI が患者転帰を改善すること。まれな事象に対して安全でないこと。臨床医を置き換えること、または臨床医を上回ること。結果が農村部や高所得地域に移転できること。費用削減が確立していること。

主な限界：観察された効果より大きな効果を検出できるよう設計されていた（まれな転帰には非常に大きなサンプルが必要）。設定エラーによって対照群が汚染され、共有ネットワーク内の習慣によって比較がぼやけた。いずれも差がない方向へのバイアスとなる。もともと高い基準で運営されていた単一の都市部民間ネットワークであること。事前に規定された非劣性または安全性の枠組みがないこと。短い 14 日間の追跡期間。

一般の読者はどの程度の確信を持つべきか。この試験でツールが安全性シグナルを示さず、記録を改善したという点には高い確信を持てる。この試験で 14 日間の患者転帰を明らかに改善しなかったという点にも高い確信を持てる。患者に対する利益という介入として「効く」または「失敗する」と主張することへの確信は低い。その問いは本当に未解決であり、はるかに大規模な研究を必要としている。適切な立場は、安全性シグナルを示さず、ケアのプロセスを改善したツールについての、慎重な現実世界の結果であって、AI がプライマリケアを変革する、あるいは破壊するという判定ではない。

出典

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