The Clean Paper

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

一個臨床 AI 看起來安全且改善了書面流程——但並未改善患者結局

一項在 *16* 家肯亞初級醫療設施中進行的實用性集群隨機試驗，測試了一個嵌入臨床官員所用電子病歷的生成式 *AI* 決策支持工具。在 *9,691* 名患者中，專家裁定的 *14* 天嚴格治療失敗複合指標在使用該工具時為 *2.2%*，未使用時為 *2.0%*（調整後比值比 *0.77*，*95% CI 0.55-1.08*，*P = 0.13*）——無顯著差異。該工具未顯示安全信號，改善了所有評分領域的文件品質，使處方和患者滿意度不變，並呈抗生素成本下降趨勢。這不是一項失敗的研究；它是一項罕見而誠實的研究——用患者而非基準測試來衡量——結果指向一個不光彩的事實：一個看起來安全且幫助了流程的工具，在兩週內未能明顯幫助患者，而要檢測任何此類益處將需要一項大得多的試驗。

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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

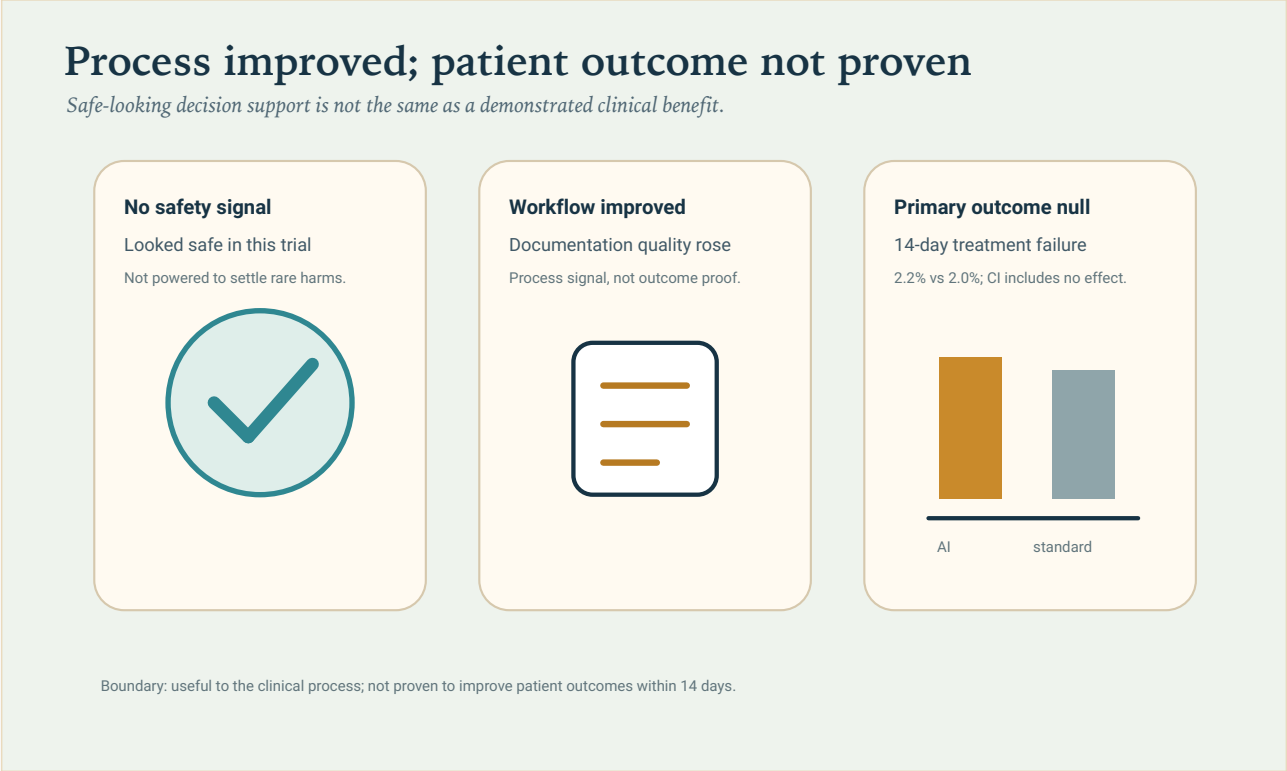
安全且有用，不等於有效

大多數關於醫療 AI 的標題都建立在錯誤類型的研究之上。一個模型在考試題目上得分很高，或者在精心策劃的病例集上擊敗了醫生，故事便不寫自成：機器已經為臨床做好了準備。

這項試驗做了更困難也更罕見的事。它將一個生成式 AI 決策支持工具放入了真實的初級醫療中，在真實的設施中，面對真實的患者，並提出了那個真正重要的問題：患者是否獲得了更好的結果？

誠實的答案是否定的——在 14 天內，沒有可測量的改善。該工具沒有顯示出安全信號。它提高了臨床文件的品質。它甚至可能降低了一些藥物成本。但它並未顯著減少治療失敗，作者謹慎地表示，任何益處如果存在，可能都是適度的。

這並不是研究的失敗。這是研究在起作用。當醫療 AI 的證據是在患者身上而非基準測試上衡量時，負責任的證據就是這個樣子的。



該工具未顯示安全信號並改善了臨床文件，但預先指定的 14 天患者結局並未顯著改善。流程幫助與經證實的患者獲益不是同一回事。

作者做了什麼

團隊在肯亞奈洛比和 Kiambu 縣一家私人醫療網絡（Penda Health）營運的 **16 家初級醫療設施**中進行了一項實用性、集群隨機試驗。這些設施中的醫療主要由**臨床官員**——擁有三年文憑的中級從業

者——提供，通常無法方便地獲得高級會診。

隨機化的單位是臨床醫生，而非患者。**103 名臨床官員**被隨機分配：52 名進入干預組，51 名進入對照組。兩組使用同一套基於雲端的電子病歷。干預組額外擁有「**AI Consult**」（2.0 版），一個基於 **OpenAI 的 GPT-4o** 大語言模型構建並嵌入該病歷的決策支持工具。它讀取臨床醫生記錄的資訊，並可標記診斷或治療計劃中可能存在的問題。臨床醫生保留完全自主權：他們可以接受、修改或忽略其建議。

> **是哪個模型，為什麼這些細節很重要**
>
> 該工具是 **AI Consult 2.0**，運行 **OpenAI 的 GPT-4o**（2025 年 5 月發布），透過 OpenAI 的商業 API 在企業授權下存取，並以低隨機性設定（溫度 0.1）運行。它位於一個客製電子病歷（EasyClinic 的 EMR）內部，並由旨在與肯亞國家治療指南保持一致的系統提示詞引導；作者發表了完整的指令提示詞。
> 為什麼要詳細說明這些？因為結果是關於一個特定系統——一個模型版本、一個提示詞、一個病歷系統、一個環境——而非關於一般意義上的「醫學中的大語言模型」。作者也提出了同樣的觀點：他們將其發現稱為一個時間基準而非能力的固定估計。一個更新的模型、一個不同的提示詞，或一個資訊化程度較低的診所，都可能改變結果。
> 關於獨立性：OpenAI 後來提供了實物支持（雲端運算積分和關於使用其 API 的技術指導），但作者聲明，使用 OpenAI 的決定是在該提議之前做出的，並且 OpenAI 在試驗設計、資料收集、分析或發表決定中沒有扮演任何角色。

在 2025 年 4 月 22 日至 7 月 16 日期間，**9,691 名患者**被納入。**主要結局被刻意設計為以患者為中心的嚴格指標**：就診後 14 天內由專家裁定的治療失敗複合指標——一個臨床醫生小組在不知曉研究分組的情況下，判斷每位患者是否出現了不良結局，如未緩解或惡化的疾病。該試驗已預先註冊（泛非臨床試驗註冊中心 202502499779176）。

這一設計選擇正是關鍵所在。展示一個 AI 工具改變了臨床醫生寫下的東西是容易的。展示它改變了患者身上發生的事情則困難得多，也重要得多。

他們發現了什麼

主要結局並未改善。治療失敗發生在 AI 組 4,693 名患者中的 102 人（2.2%）和對照組 4,654 名患者中的 94 人（2.0%）。原始百分比在 AI 輔助下略微偏高，但調整後點估計偏向益處：**調整後的比值比（aOR）為 0.77（95% 信賴區間 0.55 至 1.08，P = 0.13）**——無統計顯著性。原始數值與調整後數值之間的翻轉並非計算錯誤；調整考慮了臨床醫生集群之間的差異。無論哪種情況，信賴區間都舒適地包含「無效應」，因此無法聲稱任何益處，且以絕對數值來看，效應是微小的。

關於如何理解比值比、信賴區間和 P 值的簡明解釋，請參閱臨床結果閱讀指南。

無安全信號——在限度之內。沒有嚴重不良事件被判定與工具有關，獨立審查未發現安全信號。作者對這一保證的上限是誠實的：該試驗沒有足夠的統計檢定力來檢測罕見的嚴重傷害，也沒有預先指定的非劣效性或正式安全框架，因此它不能證明不常見事件的安全性。

文件品質得到了改善。在盲法專家審查的 2,000 次就診中，使用 AI Consult 的臨床醫生在所有評分的領域——記錄的診斷、治療計劃和整體完整性——都產生了更好的臨床文件。

處方幾乎沒有變化。在處方方面，包括正確的抗生素使用，沒有顯著差異（aOR 0.86，95% CI 0.48 至 1.55）。該工具並未改變抗生素處方率。

患者沒有注意到差異。在完成滿意度調查的 826 名患者中，兩組之間的滿意度基本相同，就診時間也相似。

成本略微向下指向。在一項調整分析中，AI 組中抗生素相關成本更低——可能是透過更便宜的選擇而非更少的處方——並且每位患者的抗生素節省似乎超過了運行該工具的每位患者成本。作者將此標記為提示性的，而非確定的：完整的總體擁有成本核算超出了試驗範圍。

作者自己的一行總結是最乾淨的版本：大語言模型輔助在這些限度內是安全的，但並未減少 14 天內的治療失敗，任何益處可能都是適度的。

為什麼「無顯著差異」不等於「它不起作用」

一個無效的主要結局很容易在任一方向上被過度解讀。有兩件事阻止了簡單的故事。

首先，**該試驗是為捕捉比其所發現的更大的效應而構建的**。初級醫療中嚴重的不良結局很罕見——此處約為 2%——因此將一個小型真實益處與噪聲區分開需要極大的樣本量。作者自己的事後統計檢定力計算表明，要檢測他們觀察到的那種大小的效應，需要一個**大得多的試驗，大約需要超過 100,000 名患者**。在這種規模的研究中，一個非顯著結果並不排除一個小型真實益處的存在；它意味著本研究無法解決這個問題。

其次，**對照組被部分模糊了**。一個組態錯誤曾短暫地讓一些對照組臨床醫生也能存取 AI Consult，而共享網絡中的臨床醫生會相互交流，並將習慣帶過邊界。這兩種效應都傾向於使兩組看起來更相似，將任何真實差異推向零。此外，宿主網絡的運行標準已經相對較高，這給工具展示改善留下了更少空間。

這些都不是在援救「突破」標題。但它們確實意味著正確的解讀是經過校準的，而非貶低的：在最困難和最誠實的終點上，該工具在兩週內未明顯幫助患者——但它確實在可測量的範圍內幫助了病歷記錄，並且看起來是安全的。

這篇論文沒有證明什麼

- 它**沒有**表明 AI 改善了患者結局。在主要 14 天終點上，沒有顯著益處。
- 它**沒有**表明 AI 是無用的。點估計傾向於它，文件全面改善，藥物成本趨勢向下；無效結果與一個小型真實益處一致，只是試驗規模太小無法確認。
- 它**沒有**證明該工具對罕見傷害是安全的。它顯示了無安全信號，但其統計檢定力或設計不足以認證不常見嚴重事件的安全性。
- 它**沒有**表明「AI 擊敗醫生」或取代臨床醫生。這是決策支持；臨床醫生保留了接受或拒絕的全部

權力。
- 它**沒有**自動泛化。該試驗在肯亞單一城市私人網絡中運行；農村、城郊和高收入環境可能在任一方向上有所不同。
- 它**沒有**確立成本節約。成本信號是提示性的，而非一項完整的經濟學評估。

證據有多強？

對於核心主張——未證明短期患者傷害的減少——證據作為設計是強有力的，作為結論則是適當謙遜的。一項前瞻性、預先註冊、集群隨機化、採用盲法、以患者為單位的複合結局試驗，接近你為這類工具能收集到的真實世界最佳證據。它比一個基準分數或一個病例集研究資訊量大得多。

對於次要發現——更好的文件、不變的處方、相似的滿意度、更低的抗生素成本——證據是良好的，但應作為次要發現來解讀：支持性信號，而非標題，且易受相同污染和單一網絡侷限性的影響。

對於安全性，證據是令人安心的但有限度的：未發現信號，但這不是一項旨在發現罕見傷害的研究。

最有用的立場既不是「它有效」也不是「它失敗了」。而是：一項謹慎的真實世界試驗發現，這個 AI 工具未產生安全信號並改善了醫療流程，但未能在兩週內展示患者結局的益處——而要檢測任何此類益處將需要一項大得多的研究。

為什麼重要

關於醫療 AI 的辯論正缺乏的正是這類證據。有成千上萬的論文展示模型在考試中取得優異成績並在整潔病例上與臨床醫生匹敵。但很少有大型、實用性、隨機化試驗衡量真實患者是否做得更好。這是其中之一，它落在了不光彩的真相上：通過考試與幫助患者並不是一回事。

這一差距就是整個故事。一個工具可以對臨床醫生真正有用——更清晰的筆記、更低的藥費、第二雙眼睛——但仍然無法在兩週內推動一個硬性患者結局。這兩個事實可以同時成立，而一個成熟的醫療系統必須將它們同時容納，而非選擇方便的那個。

它還將舉證責任重置到一個有用的方向。如果一家公司想聲稱其臨床 AI 改善了診療，相關的證據不是一個排行榜。而是一項像這樣的試驗，衡量對患者重要的結局——而且最好是一項更大的試驗，因為這裡的誠實教訓是，適度的益處需要大規模的研究才能看到。

清潔摘要

一項在 16 家肯亞初級醫療設施中進行的實用性集群隨機試驗，測試了一個嵌入臨床官員所用電子病歷的生成式 AI 決策支持工具（「AI Consult」）。在 9,691 名患者中，專家裁定的 14 天內治療失敗複合指標在使用該工具時為 2.2%，未使用時為 2.0%（調整後比值比 0.77，95% CI 0.55–1.08，P = 0.13）——無顯著差異。該工具未顯示安全信號，改善了所有評分領域的臨床文件品質，未改變處方，使患

者滿意度不變，並與略低的抗生素成本相關。作者得出結論，它在這些限度內是安全的，但並未減少治療失敗，任何益處可能都是適度的；要檢測所觀察到的那種大小的效應，將需要一項大得多的試驗（約 100,000 名患者規模）。該結果並未表明臨床 AI 能改善患者結局，也未表明它是無用的——它表明，一個看起來安全並幫助了流程的工具，在兩週內未能在單一城市私人網絡中明顯幫助患者。

No-BS 檢查

論文所展示的：在一項真實世界隨機試驗中，將一個基於大語言模型的決策支持工具添加到初級醫療病歷中未產生安全信號，改善了臨床文件品質，未改變處方，且未顯著減少一項嚴格的 14 天患者治療失敗複合指標。

合理但未經證實的：該工具產生了試驗規模太小而無法檢測的小型真實治療失敗減少；一旦計入全部成本它能節省資金；更好的文件最終能轉化為更好的診療。

它沒有展示的：臨床 AI 能改善患者結局；它對罕見事件是安全的；它取代或超越了臨床醫生；這些結果可轉移至農村或高收入環境；成本節約已經確立。

主要侷限：統計檢定力針對比所觀察到更大的效應量（罕見結局需要極大樣本）；一個組態錯誤污染了對照組，共享網絡中的習慣模糊了對比，兩者都偏向於無差異；單一城市私人網絡且標準已經較高；無預先指定的非劣效性或安全框架；僅限 14 天短期視野。

普通讀者應該抱有多大信心？高度信心——該工具在本試驗中未顯示安全信號且改善了文件。高度信心——它在此處未明顯改善 14 天患者結局。對於任何聲稱它作為一項患者獲益干預「有效」或「失敗」的說法，信心應低——這一問題確實尚未解決，需要一項大得多的研究。恰當的立場：一個關於一項工具的安全且改善了醫療流程的謹慎真實世界結果——而非一項判決 AI 能改變或毀掉初級醫療的裁決。

來源

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