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群集隨機試驗，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——結果指向一個不那麼吸睛但重要的結論：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 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

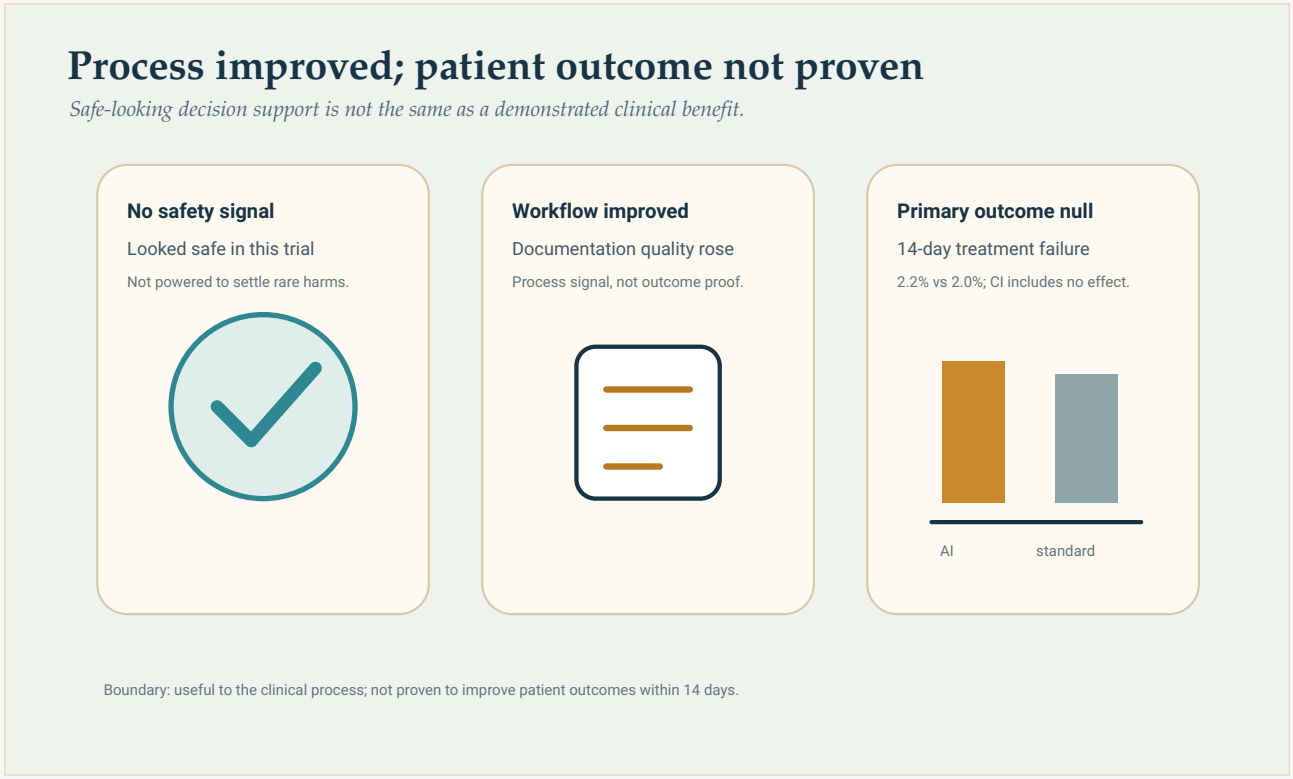
安全且有幫助，不等於有效

多數醫療 AI 頭條，都是從類型不對的研究拼湊出來的。模型在考試題目上得分很高，或在精心挑選的病例情境中勝過醫師，故事便自然成形：機器已經準備好進入臨床。

這項試驗做了更困難、也更少見的事。它把生成式 AI 決策支援工具放進真實的基層醫療，在真實的醫療機構面對真實的患者，並提出真正重要的問題：患者是否因此得到更好的結果？

誠實的答案是否定的——至少在 14 天內，沒有可測量的改善。這項工具沒有顯示安全性訊號，提升了臨床文件的品質，甚至可能降低部分藥物成本。但它沒有顯著減少治療失敗，而作者也謹慎表示，任何效益如果確實存在，可能都很有限。

這不是研究失敗了，而是研究發揮了作用。當 AI 在患者身上、而不是在基準測試上接受評估時，負責任的醫療 AI 證據就是這個樣子。



這項工具沒有顯示安全性訊號，並改善了臨床文件，但預先指定的 14 天患者結果並未顯著改善。改善流程不等於已證實的患者效益。

作者做了什麼

研究團隊在肯亞奈洛比郡和基安布郡的 **16 所基層醫療機構**中，進行了一項務實的叢集隨機試驗；這些機構由私人醫療網路 Penda Health 營運。這些機構的醫療照護主要由 **臨床官**提供——他們是持有

三年制文憑的中階醫療人員，通常不容易取得資深醫療人員的會診。

隨機分派的單位是臨床人員，而不是患者。共有 **103 名臨床官**被隨機分派：52 人進入介入組，51 人進入對照組。兩組都使用相同的雲端電子病歷。介入組另外使用 「**AI Consult**」（版本 2.0），這是一項建置於 **OpenAI 的 GPT-4o** 大型語言模型之上、並嵌入該病歷的決策支援工具。它會讀取臨床人員記錄的資訊，並可能提示診斷或治療計畫中的問題。臨床人員保有完整自主權：可以接受、修改或忽略它的建議。

> **使用的是哪個模型，以及細節為何重要**
>
> 這項工具是 **AI Consult 2.0**，執行於 **OpenAI 的 GPT-4o**（2025 年 5 月發布版本），透過企業授權下的 OpenAI 商業 API 使用，並採用低隨機性設定（temperature 0.1）。它位於客製化電子病歷 EasyClinic 的 EMR 中，並由依照肯亞國家治療指引撰寫的系統提示詞引導；作者公開了完整的指令提示詞。
>
> 為什麼要交代這些細節？因為這項結果談的是一個特定系統——一個模型版本、一組提示詞、一套病歷、一種設定——而不是泛指「醫學中的 LLM」。作者也提出同一點：他們稱這項發現是時間基準，而非能力的固定估計。更新的模型、不同的提示詞，或數位化程度較低的診所，都可能改變結果。
>
> 關於獨立性：OpenAI 後來提供了實物形式的支援（雲端運算額度，以及使用其 API 的技術指導），但作者表示，採用 OpenAI 的決定是在這項提議之前作出的，而且 OpenAI 沒有參與試驗設計、資料收集、分析或發表決定。

2025 年 4 月 22 日至 7 月 16 日期間，共有 **9,691 名患者**入組。**主要結果刻意以患者為中心且標準嚴格**：就診後 14 天內、由專家判定的治療失敗複合結果——由一組對研究分組不知情的臨床人員判定每位患者是否出現疾病未緩解或惡化等不良結果。這項試驗事先完成註冊（泛非臨床試驗登錄庫 202502499779176）。

這項設計選擇正是重點。要證明 AI 工具改變了臨床人員記錄的內容很容易；要證明它改變了患者身上發生的事，則困難得多，也有意義得多。

他們發現了什麼

主要結果沒有改善。 AI 組 4,693 名患者中有 102 人（2.2%）出現治療失敗，對照組 4,654 人中有 94 人（2.0%）出現治療失敗。AI 組的原始百分比略高，但調整後的點估計傾向於有益：**調整後勝算比為 0.77（95% 信賴區間 0.55 至 1.08，P = 0.13）**——不具統計顯著性。原始數字與調整後數字之間的這種反轉不是算術疏漏；調整反映了不同臨床人員叢集之間的差異。無論如何，信賴區間明確包含「沒有影響」的可能，因此不能宣稱有益，而且以絕對值來看，效果非常小。

若想用白話了解如何一起解讀勝算比、信賴區間和 P 值，請參閱臨床結果指南。

沒有安全性訊號——但有其限制。沒有任何嚴重不良事件被判定與這項工具有關，獨立審查也沒有發現安全性訊號。作者誠實說明了這份安心感的上限：這項試驗沒有足夠統計效力偵測罕見的嚴重傷害，也沒有預先指定非劣性或正式的安全性架構，因此無法證明罕見事件具安全性。

文件品質變好了。在由不知道分組的專家審查的 2,000 次就診中，使用 AI Consult 的臨床人員在所有評分面向上都產出較好的臨床文件——包括記錄的診斷、治療計畫和整體完整性。

處方幾乎沒有變化。處方方面沒有顯著差異，包括正確使用抗生素（調整後勝算比 0.86，95% CI 0.48 至 1.55）。這項工具沒有改變抗生素處方率。

患者沒有察覺差異。在完成滿意度問卷的 826 名患者中，兩組的滿意度幾乎相同，問診時間也相近。

成本略微下降。在一項調整後分析中，AI 組與抗生素相關的成本較低——可能是因為選用了較便宜的藥物，而不是因為處方數量減少——而且每位患者節省的抗生素費用似乎高於每位患者執行這項工具的成本。作者指出，這只是具有提示性的結果，尚未定論：完整的總持有成本核算不在本試驗範圍內。

作者自己的單句總結最清楚：在這些限制之內，LLM 輔助是安全的，但沒有在 14 天內降低治療失敗，任何效益可能都很有限。

為什麼「沒有顯著差異」不等於「它沒用」

主要結果為無效結果，很容易在任一方向上被過度解讀。有兩件事阻止我們採用過於簡單的說法。

第一，**這項試驗原本是為了捕捉比實際發現更大的效果而設計的**。基層醫療中的嚴重不良結果很少見——這裡約為 2%——因此要把小幅的真實效益與雜訊區分開來，需要極大的樣本數。作者自己事後進行的統計效力計算顯示，要偵測到他們觀察到的幅度，可能需要一項**大得多的試驗，規模約超過 100,000 名患者**。這種規模的研究得到不顯著結果，並不能排除小幅但真實的效益；它表示這項研究無法辨別這種效益。

第二，**兩組之間的比較有部分模糊**。一項設定錯誤曾短暫讓一些對照組臨床人員取得 AI Consult 的使用權，而同一醫療網路中的臨床人員會彼此交流，也會把習慣帶過分組界線。這兩種影響都傾向讓兩組看起來更相似，使任何真實差異朝零靠近。此外，該醫療網路原本就已維持相對高的標準，讓工具能展現改善的空間更小。

這些都不足以挽救「突破性進展」的頭條。但這確實表示，正確的解讀應該是經過校準，而不是唱衰：在最困難、也最誠實的終點上，這項工具沒有證明能在兩週內幫助患者——同時，它確實可測量地改善了紀錄工作，看起來也很安全。

這項研究沒有證明什麼

- 它**沒有**顯示 AI 改善了患者結果。在主要的 14 天終點上，沒有顯著效益。
- 它**沒有**顯示 AI 毫無用處。點估計偏向 AI，文件品質全面改善，藥物成本也呈下降趨勢；無效結果與小幅但真實、只是試驗規模太小而無法確認的效益相容。
- 它**沒有**證明這項工具對罕見傷害安全。它沒有顯示安全性訊號，但試驗沒有足夠統計效力，也不是為了替罕見嚴重事件的安全性作認證而設計。
- 它**沒有**顯示「AI 勝過醫師」或取代臨床人員。這是決策支援；臨床人員保有接受或拒絕建議的完

整權限。
- 它**不會自動**適用於其他情境。試驗在肯亞一個城市私人網路中進行；鄉村、城市周邊和高收入環境可能朝任一方向不同。
- 它**沒有**確立成本節省。成本訊號具有提示性，並不是完整的經濟評估。

證據有多強？

對於核心主張——沒有已證實的短期患者傷害降低——這份證據就設計而言很強，就結論而言則保持了適當的謙遜。一項前瞻性、預先註冊、叢集隨機分派的試驗，搭配由不知道分組的審查者判定、以患者為層級的複合結果，已接近能為這類工具收集到的最佳真實世界證據。它提供的資訊遠勝於基準測試分數或病例情境研究。

對於次要發現——文件品質更好、處方沒有變化、滿意度相近、抗生素成本較低——證據良好，但應將其視為次要結果：它們是支持性的訊號，不是主要焦點，而且容易受到相同的污染問題和單一醫療網路限制影響。

至於安全性，證據令人安心，但有明確界線：沒有發現訊號，卻不是一項為了找出罕見傷害而設計的研究。

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，這項 AI 工具沒有顯示安全性訊號，並改善了照護流程，卻沒有在兩週內證明能改善患者結果——而要偵測任何這類效益，需要大得多的研究。

為什麼重要

醫療 AI 的討論正缺少這種證據。有數千篇論文顯示模型在考試中表現優異，並在精心整理的病例上與臨床人員相當。但真正衡量患者是否得到更好結果的大型務實隨機試驗寥寥可數。這項研究就是其中之一，而它落在一個不吸引人的真相上：通過測驗不等於幫助患者。

這個落差就是整個故事。一項工具可以確實對臨床人員有用——筆記更清楚、藥費更低、多一雙審視的眼睛——卻仍然無法在兩週內改變一項嚴格的患者結果。兩個事實可以同時成立，而成熟的醫療體系必須把它們放在一起理解，而不是挑選方便的那一個。

這也把舉證責任重新導向一個有益的方向。如果公司想宣稱其臨床 AI 能改善照護，相關證據不是排行榜，而是這樣的試驗：以患者在意的結果為衡量標準——而且理想上規模要更大，因為這裡的誠實教訓是，適度的效益需要大型研究才能看見。

重點摘要

一項在肯亞 16 所基層醫療機構進行的務實叢集隨機試驗，測試了加入臨床官所使用電子病歷的生成式 AI 決策支援工具（「AI Consult」）。在 9,691 名患者中，14 天內由專家判定的治療失敗複合結果，在使用工具時為 2.2%，未使用時為 2.0%（調整後勝算比 0.77，95% CI 0.55–1.08，P = 0.13）——沒有顯著差異。這項工具沒有顯示安全性訊號，改善了所有受評面向的臨床文件，沒有改變處方，患者滿意度沒有變化，且與略低的抗生素成本相關。作者總結，在這些限制內它是安全的，但沒有降低治療失敗，任何效益可能都很有限；要偵測觀察到的幅度，將需要大得多的試驗（規模約 100,000 名患者）。這項結果沒有顯示臨床 AI 能改善患者結果，也沒有顯示它毫無用處——它顯示的是：一項看起來安全且能改善流程的工具，在一個城市私人網路中，沒有證明能在兩週內幫助患者。

務實檢視

論文顯示的內容：在一項真實世界隨機試驗中，將以 LLM 為基礎的決策支援工具加入基層醫療紀錄，未發現安全性訊號，改善了臨床文件品質，沒有改變處方，也沒有顯著降低嚴格定義的 14 天患者治療失敗複合結果。

合理但尚未證明的內容：這項工具能讓治療失敗小幅但真實地減少，只是幅度小到本試驗無法偵測；計入全部成本後，它能節省金錢；更好的文件最終會轉化為更好的照護。

它沒有顯示的內容：臨床 AI 能改善患者結果；它對罕見事件不安全；它能取代或勝過臨床人員；這些結果能延伸到鄉村或高收入環境；成本節省已獲確立。

主要限制：試驗的統計效力是按大於實際觀察到的效果設計的（罕見結果需要非常大的樣本）；設定錯誤污染了對照組，加上同一網路的習慣使比較變得模糊，兩者都會使結果偏向沒有差異；單一、原本標準已很高的城市私人網路；沒有預先指定的非劣性或安全性架構；14 天的觀察期很短。

一般讀者應有多大信心？對於這項工具在本試驗中沒有顯示安全性訊號並改善文件品質，可以有高度信心。對於它在這裡沒有證明能改善 14 天患者結果，也可以有高度信心。至於它作為以患者獲益為目標的介入措施「有效」或「失敗」的任何說法，信心應很低——這個問題確實仍未解決，需要大得多的研究。適當的立場是：這是一項關於某項未發現安全性訊號、並改善照護流程之工具的謹慎真實世界結果，而不是宣判 AI 能改造——或摧毀——基層醫療。

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