The Clean Paper

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

一个临床 AI 看起来安全且改善了书面流程——但并未改善患者结局

一项在 *16* 家肯尼亚初级医疗设施中进行的实用性集群随机试验，测试了一个嵌入临床官员所用电子病历的生成式 *AI* 决策支持工具。在 *9,691* 名患者中，专家裁定的 *14* 天严格治疗失败复合指标在使用该工具时为 *2.2%*，未使用时为 *2.0%*（调整后比值比 *0.77*，*95% CI 0.55-1.08*，*P = 0.13*）——无显著差异。该工具未显示安全信号，改善了所有评分领域的文档质量，使处方和患者满意度不变，并呈抗生素成本下降趋势。这不是一项失败的研究；它是一项罕见而诚实的研究——用患者而非基准测试来衡量——结果指向一个不那么吸睛但重要的结论：一个看起来安全且帮助了流程的工具，在两周内未能明显帮助患者，而要检测任何此类益处将需要一项大得多的试验。

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

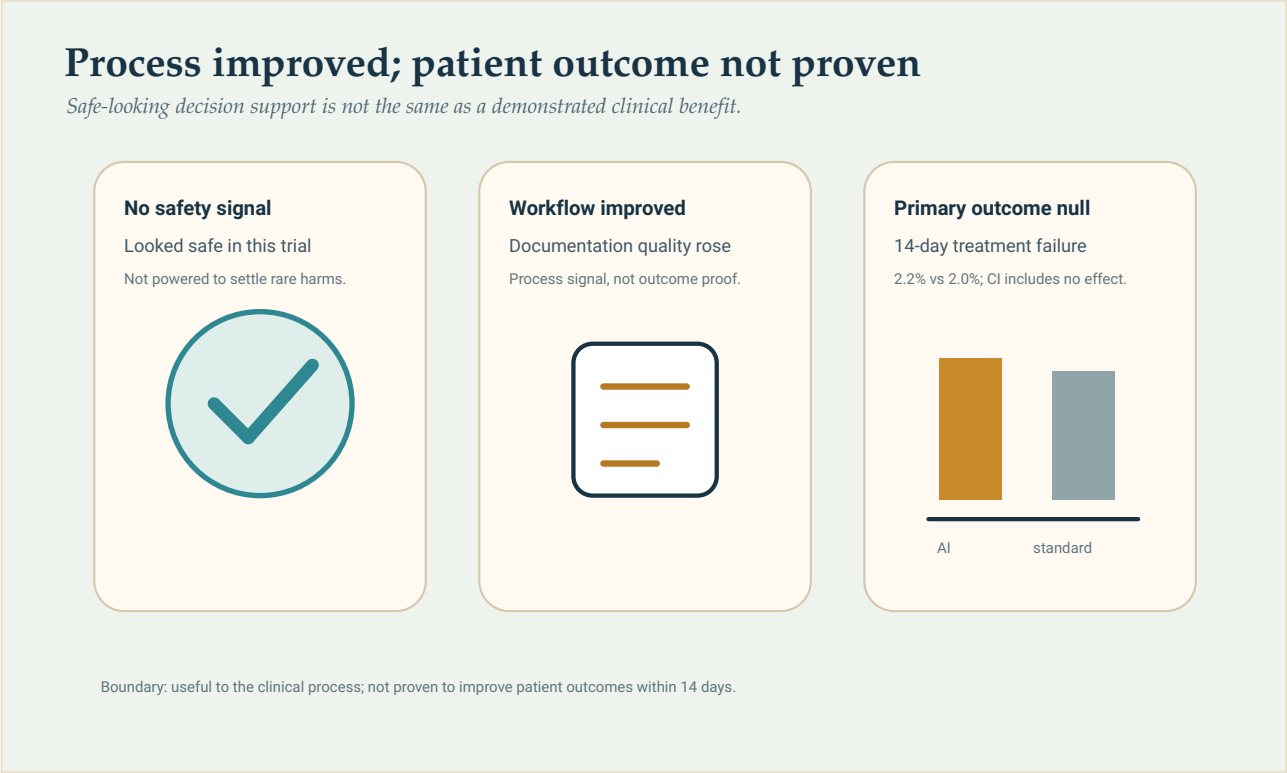
安全且有帮助，不等于有效

大多数关于医疗 AI 的头条新闻，都建立在错误类型的研究之上。模型在考试题上得分不错，或者在精选病例摘要上击败医生，于是故事便顺理成章：这台机器已经可以进入诊所了。

这项试验做了一件更难、也更少见的事。它把一个生成式 AI 决策支持工具置于真实的初级卫生保健环境中，部署在真实的医疗机构，面对真实的患者，并提出了真正重要的问题：患者的情况是否变好了？

诚实的答案是否定的——至少在 14 天内，没有可测量的改善。该工具没有显示出安全性信号。它提高了临床记录的质量，甚至可能降低了部分药物成本。但它没有显著减少治疗失败，作者也谨慎地表示，任何可能存在的获益都很可能是有限的。

这并不是研究失败了，而是研究发挥了作用。当 AI 不再根据基准测试，而是根据患者来评估时，负责任的医学 AI 证据就应当是这个样子。



该工具没有显示出安全性信号，并改善了临床记录，但预先设定的 14 天患者结局没有显著改善。对流程的帮助不等于已被证实的患者获益。

作者做了什么

研究团队在肯尼亚内罗毕和基阿姆布县一家私立医疗网络（Penda Health）运营的 **16 家初级卫生保健机构**中，开展了一项务实性、整群随机试验。这些机构的医疗服务主要由 **临床官员** 提供——他们是

持有三年制文凭的中级医疗从业者，通常不容易获得高级医生的咨询。

随机化单位是临床医生，而不是患者。共有 **103 名临床官员** 被随机分组：52 人进入干预组，51 人进入对照组。两组使用同一个基于云的电子病历系统。干预组另外获得了 **“AI Consult”**（2.0 版），这是一个构建于 **OpenAI 的 GPT-4o** 大型语言模型之上、并嵌入该病历系统的决策支持工具。它会读取临床医生记录的信息，并提示诊断或治疗方案中可能存在的问题。临床医生始终拥有完全自主权：可以接受、修改或忽略它的建议。

> **使用的是哪个模型，为什么这些细节很重要**
>
> 该工具是 **AI Consult 2.0**，运行于 **OpenAI 的 GPT-4o**（2025 年 5 月发布），通过企业许可下 OpenAI 的商业 API 接入，并采用低随机性参数（temperature 0.1）。它位于定制的电子病历系统（EasyClinic 的 EMR）中，并由按照肯尼亚国家治疗指南编写的系统提示词引导；作者公开了完整的指令提示词。
> 为什么要把这些说清楚？因为这个结果针对的是 一个具体的系统——一个模型版本、一套提示词、一种病历系统、一个使用环境——而不是泛泛而谈的 “医学中的 LLM”。作者也强调了同一点：他们把这一发现称为 时间基准，而不是能力的固定估计。更新的模型、不同的提示词，或者数字化程度较低的诊所，都可能改变结果。
> 关于独立性：OpenAI 后来提供了实物支持（云计算额度，以及如何使用其 API 的技术指导），但作者表示，决定使用 OpenAI 是在这项支持提出 之前作出的，而且 OpenAI 没有参与试验设计、数据收集、分析或发表决定。

2025 年 4 月 22 日至 7 月 16 日期间，共有 **9,691 名患者**入组。**主要结局被有意设定为以患者为中心且标准严格**：就诊后 14 天内由专家判定的 治疗失败复合结局——由一组不知道研究分组的临床医生判断每位患者是否出现了未解决或加重的病情等不良结局。该试验已提前注册（泛非临床试验注册库 202502499779176）。

这项设计选择正是关键。证明 AI 工具改变了临床医生记录的内容很容易；证明它改变了患者身上发生的事情，则困难得多，也更有意义。

他们发现了什么

主要结局没有改善。 AI 组 4,693 名患者中有 102 人发生治疗失败（2.2%），对照组 4,654 人中有 94 人发生治疗失败（2.0%）。未经校正的百分比在使用 AI 时略高，但校正后点估计转而倾向于获益：**校正后的比值比为 0.77（95% 置信区间 0.55 至 1.08，P = 0.13）**——没有统计学显著性。原始数字和校正数字之间的这种反转不是算术错误；校正考虑了临床医生群组之间的差异。无论如何，置信区间都充分包含 “无效应” 的可能，因此不能宣称存在获益；而且从绝对值看，效应非常小。

如果想用通俗英语一起理解比值比、置信区间和 P 值，请参阅临床结果指南。

没有安全性信号——但存在局限。没有严重不良事件被判定与该工具有关，独立审查也没有发现安全性信号。作者坦率地说明了这种令人安心的结论的边界：该试验没有足够的统计效能来检测罕见的严重伤害，也没有预先设定的非劣效性或正式安全性框架，因此无法 证明该工具对不常见事件是安全的。

记录质量提高了。在由不知道分组情况的专家审查的 2,000 次就诊中，使用 AI Consult 的临床医生在所有评估领域都产出了更好的临床记录——包括记录的诊断、治疗方案和总体完整性。

处方几乎没有变化。处方方面没有显著差异，包括抗生素的正确使用（校正后的比值比为 0.86，95% CI 0.48 至 1.55）。该工具没有改变抗生素处方率。

患者没有察觉到差异。在完成满意度调查的 826 名患者中，两组满意度基本相同，问诊时间也相近。

成本略有下降。在一项校正分析中，AI 组与抗生素相关的成本较低——可能是因为选择了更便宜的药物，而不是因为处方数量减少——而且每位患者节省的抗生素费用似乎超过了运行该工具的每位患者成本。作者指出，这只是提示性结果，尚未定论：完整的总拥有成本核算不在本试验范围内。

作者自己的单句总结最为简洁：在这些限制范围内，LLM 辅助是安全的，但没有在 14 天内减少治疗失败，任何获益都很可能是有限的。

为什么 “没有显著差异” 不等于 “它不起作用”

无效的主要结局很容易被往任一方向过度解读。有两点阻止我们接受过于简单的说法。

第一，**这项试验原本是为捕捉比实际发现更大的效应而设计的**。初级卫生保健中的严重不良结局很少见——本研究中约为 2%——因此，要从噪声中分辨出一个较小但真实的获益，需要极大的样本量。作者自己事后进行的统计效能计算表明，要检测到他们观察到的效应大小，可能需要一项**规模大得多的试验，患者数量超过 100,000 人**。如此规模的研究得到非显著结果，并不能排除一个小而真实的获益；它只能说明这项研究无法确认这一点。

第二，**两组之间的比较部分被模糊了**。一次配置错误曾让一些对照组临床医生短暂获得 AI Consult 的使用权限；而且同一医疗网络中的临床医生会彼此交流，并把习惯带过分组边界。这两种影响都会让两组看起来更相似，使任何真实差异趋近于零。除此之外，该医疗网络本身已经按照相对较高的标准运行，这也减少了工具显示改善的空间。

这些情况都不足以挽救 “突破性进展” 的头条新闻。但它们意味着，正确的解读应当是校准后的，而不是泄气式的：在最困难、也最诚实的终点上，该工具没有证明能在两周内帮助患者——与此同时，它确实可测量地改善了记录工作，看起来也是安全的。

这项研究没有证明什么

- 它**没有**显示 AI 改善了患者结局。在 14 天主要终点上，没有显著获益。
- 它**没有**显示 AI 毫无用处。点估计倾向于 AI，记录质量全面改善，药物成本也呈下降趋势；无效结果与一个小而真实、但试验规模不足以确认的获益相容。
- 它**没有**证明该工具对罕见伤害是安全的。它没有显示安全性信号，但试验没有足够的统计效能，也没有被设计用于确认不常见严重事件的安全性。
- 它**没有**显示 “AI 击败医生” 或取代临床医生。这是决策支持；临床医生仍完全有权接受或拒绝它。
- 它**不会**自动适用于其他环境。试验在肯尼亚一个私立城市医疗网络中进行；农村、城郊和高收入

环境可能朝任一方向有所不同。
- 它**没有**确立成本节省。成本信号只是提示性的，并不是一项完整的经济学评估。

证据有多强?

对于核心主张——没有已被证实的短期患者伤害下降——这项证据在设计层面很强，在结论层面则恰当地保持了审慎。一项前瞻性、预先注册、整群随机的试验，采用盲法、以患者为单位的复合结局，接近于针对这类工具所能收集到的最佳真实世界证据。它提供的信息远多于基准测试分数或病例摘要研究。

对于次要发现——记录质量提高、处方没有变化、满意度相近、抗生素成本下降——证据是不错的，但应当作为次要结果来理解：它们是支持性信号，而不是头条结论，并且容易受到同样的污染问题和单一医疗网络局限的影响。

对于安全性，证据令人放心，但有边界：没有发现信号，但这不是一项旨在发现罕见伤害的研究。

最有用的态度既不是 “它有效”，也不是 “它失败了”。准确的说法是：一项谨慎的真实世界试验发现，这款 AI 工具没有引发安全性信号，改善了医疗流程，却没有证明能在两周内带来患者结局获益——而要检测到任何这样的获益，需要规模大得多的研究。

为什么这很重要

医疗 AI 的讨论恰恰缺少这类证据。成千上万篇论文显示，模型能在考试中取得优异成绩，在整洁的病例上与临床医生表现相当。但真正测量真实患者是否获益的大型务实性随机试验寥寥无几。这项研究就是其中之一，并且揭示了一个不那么光鲜的事实：通过测试不等于帮助患者。

这道鸿沟就是全部故事。一款工具可以真正帮助临床医生——让记录更清楚、药费更低、多一双审视的眼睛——却仍然无法在两周内改变一个严格的患者结局。这两件事可以同时为真；成熟的卫生系统必须把它们放在一起考虑，而不是挑选那个更方便的结论。

它还以一种有益的方式重新设定了举证责任。如果一家公司想声称其临床 AI 能改善医疗服务，那么相关证据不是排行榜，而是像这项研究一样、围绕对患者重要的结局开展的试验——理想情况下还应当更大，因为这里最诚实的教训是：要看见有限的获益，需要大规模研究。

简明总结

一项在肯尼亚 16 家初级卫生保健机构开展的务实性整群随机试验，测试了一个添加到临床官员所使用电子病历中的生成式 AI 决策支持工具（“AI Consult”）。在 9,691 名患者中，专家判定的 14 天内治疗失败复合结局，在使用该工具时为 2.2%，未使用时为 2.0%（校正后的比值比 0.77，95% CI 0.55–1.08，P = 0.13）——没有显著差异。该工具没有显示出安全性信号，改善了所有被评估领域的临床记录，没有改变处方，患者满意度也没有变化，并且与抗生素成本略有下降相关。作者的结论是，在这些限制

范围内，该工具是安全的，但没有减少治疗失败，任何获益都很可能有限；要检测到所观察效应大小的影响，需要规模大得多的试验（患者数量超过 100,000 人）。这一结果没有显示临床 AI 能改善患者结局，也没有显示它毫无用处——它显示的是：在一个城市私立医疗网络中，一款看起来安全且有助于医疗流程的工具，没有证明能在两周内帮助患者。

不绕弯检查

论文显示了什么：在一项真实世界随机试验中，把基于 LLM 的决策支持工具加入初级卫生保健记录后，没有引发安全性信号，提高了临床记录质量，没有改变处方，也没有显著减少严格的 14 天患者治疗失败复合结局。

合理但尚未证实的内容：该工具可能使治疗失败真实地小幅减少，只是幅度小到本试验无法检测；在计算全部成本后，它可能节省资金；更好的记录最终可能转化为更好的医疗服务。

它没有显示什么：临床 AI 能改善患者结局；它对罕见事件不安全；它能取代临床医生或胜过临床医生；这些结果能推广到农村或高收入环境；成本节省已经确立。

主要局限：试验的统计效能是按大于实际观察到的效应来设计的（罕见结局需要非常大的样本量）；配置错误污染了对照组，而共享医疗网络中的习惯模糊了比较，这两点都会使结果偏向没有差异；研究只涉及一个本已采用较高标准的私立城市医疗网络；没有预先设定的非劣效性或安全性框架；观察期只有 14 天。

普通读者应有多大信心？可以高度确信，该工具在这项试验中没有显示出安全性信号，并改善了记录质量。可以高度确信，它在这里没有证明能改善 14 天患者结局。对于它作为一种改善患者获益的干预“有效”或“失败”的任何说法，信心都应当很低——这个问题确实尚未解决，需要规模大得多的研究。恰当的态度是：这是关于一款没有引发安全性信号、改善了医疗流程的工具的谨慎真实世界结果，而不是宣判 AI 能改变——或摧毁——初级卫生保健。

来源

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