The
CLEAN PAPER

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

Clinical AI ที่ดูปลอดภัยและทำเอกสารดีขึ้น — แต่ไม่ได้ปรับ patient outcome

Pragmatic cluster-randomized trial ใน primary care เคนยา 16 แห่งทดสอบ generative-AI decision-support tool ที่เพิ่มเข้า record ของ clinical officer ในผู้ป่วย 9,691 คน composite treatment failure 14 วันที่ expert adjudicate คือ 2.2% เมื่อมี tool เทียบ 2.0% เมื่อไม่มี (adjusted odds ratio 0.77, 95% CI 0.55-1.08, P = 0.13) — ไม่มีความแตกต่างมีนัยสำคัญ Tool ไม่แสดง safety signal ปรับ documentation ทุก domain ไม่เปลี่ยน prescribing หรือ satisfaction และมีแนวโน้มลด antibiotic cost นี่ไม่ใช่งานล้มเหลว แต่เป็นงานหายากที่ซื่อตรง — วัดบนผู้ป่วย ไม่ใช่ benchmark — และลงจอดบนความจริงว่าเครื่องมือซึ่งดูปลอดภัยและช่วย process ไม่ได้ช่วยผู้ป่วยอย่างพิสูจน์ได้ในสองสัปดาห์ และหากจะจับประโยชน์เล็กต้อง trial ใหญ่กว่ามาก

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

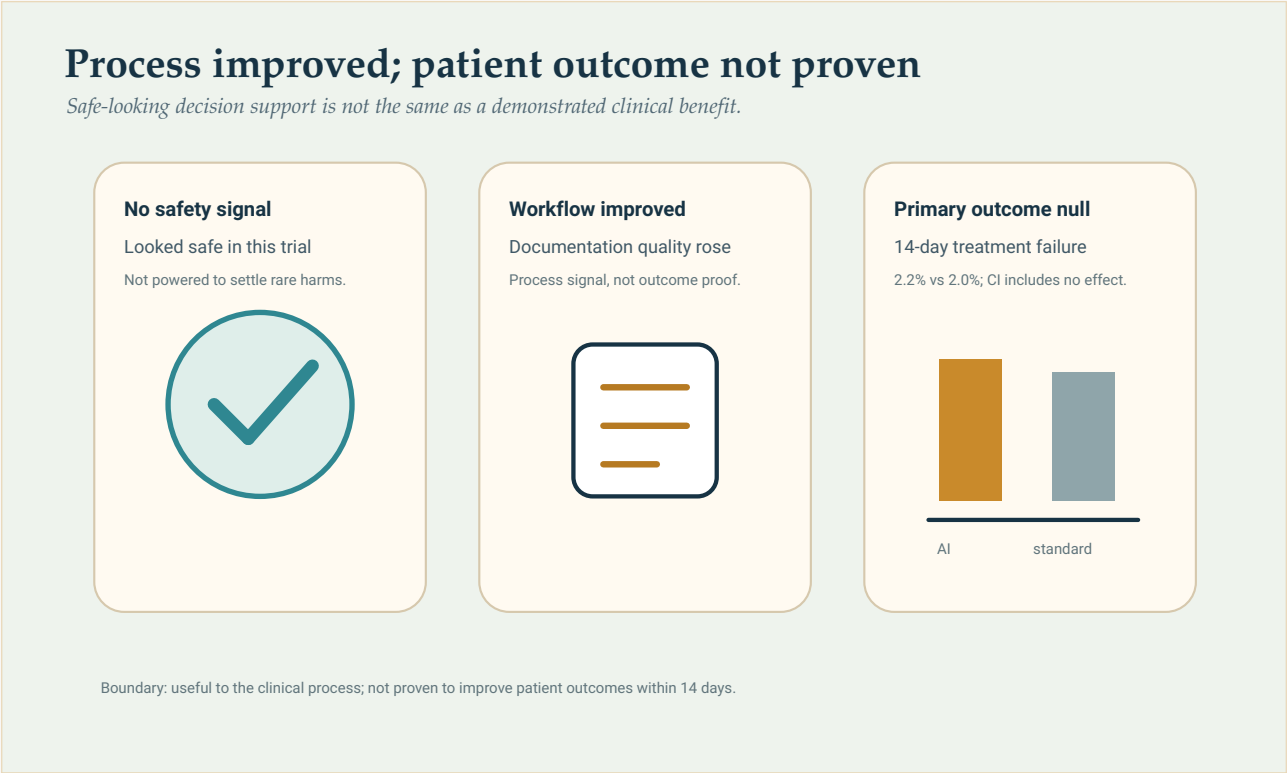
ปลอดภัยและช่วยงานได้ ไม่ได้แปลว่าช่วยผู้ป่วยได้จริง

พาดหัวเกี่ยวกับ AI ทางการแพทย์จำนวนมากสร้างจากงานศึกษาผิดประเภท โมเดลทำข้อสอบได้ดี หรือทำผลงานเหนือแพทย์ในกรณีตัวอย่างที่คัดมาอย่างเป็นระเบียบ แล้วเรื่องเล่าก็เขียนตัวเองต่อทันทีว่า เครื่องมือนี้พร้อมใช้ในคลินิกแล้ว

งานทดลองนี้ทำสิ่งที่ยากกว่าและพบได้น้อยกว่ามาก นักวิจัยนำเครื่องมือ AI เชิงกำเนิดสำหรับช่วยตัดสินใจเข้าไปใช้ในการดูแลปฐมภูมิจริง ในสถานพยาบาลจริง กับผู้ป่วยจริง แล้วถามคำถามที่สำคัญที่สุดว่า **ผู้ป่วยมีผลลัพธ์ดีขึ้นหรือไม่**

คำตอบที่ตรงที่สุดคือ **ยังไม่เห็นว่าดีขึ้น** อย่างน้อยก็ไม่สามารถตรวจพบได้ภายใน 14 วัน เครื่องมือไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจน ช่วยให้การบันทึกข้อมูลทางคลินิกดีขึ้น และอาจลดค่ายาบางส่วนด้วย แต่ไม่ได้ลดการรักษาล้มเหลวอย่างมีนัยสำคัญทางสถิติ ผู้เขียนจึงระมัดระวังที่จะสรุปว่า หากมีประโยชน์ต่อผู้ป่วยจริง ขนาดของประโยชน์ก็น่าจะไม่ใหญ่

นี่ไม่ใช่ความล้มเหลวของงานวิจัย ตรงกันข้าม นี่คืองานวิจัยที่กำลังทำหน้าที่ของมัน หลักฐานที่รับผิดชอบเกี่ยวกับ AI ทางการแพทย์ควรมีหน้าตาแบบนี้ เมื่อเราวัดผลที่เกิดกับผู้ป่วยจริง แทนที่จะดูเพียงคะแนนจากชุดทดสอบมาตรฐาน



เครื่องมือไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจน และช่วยให้การบันทึกข้อมูลทางคลินิกดีขึ้น แต่ผลลัพธ์ของผู้ป่วยภายใน 14 วันที่กำหนดไว้ล่วงหน้าไม่ได้ดีขึ้นอย่างมีนัยสำคัญ การช่วยให้กระบวนการทำงานดีขึ้นไม่เท่ากับการพิสูจน์ว่าผู้ป่วยได้ประโยชน์

The Clean Paper CC BY 4.0

ผู้วิจัยทำอะไร

ทีมวิจัยทำการทดลองเชิงปฏิบัติแบบสุ่มเป็นกลุ่มใน **สถานพยาบาลปฐมภูมิ 16 แห่ง** ของเครือข่ายสุขภาพเอกชน Penda Health ในเขต Nairobi และ Kiambu ประเทศเคนยา การดูแลในสถานพยาบาลเหล่านี้ส่วนใหญ่ดำเนินการโดย **เจ้าหน้าที่**

เวชปฏิบัติ (clinical officers) ซึ่งเป็นบุคลากรทางการแพทย์ระดับกลางที่ผ่านหลักสูตรประกาศนียบัตรสามปี และมักเข้าถึงคำปรึกษาจากแพทย์อาวุโสได้ไม่ง่าย

หน่วยที่ใช้สุ่มไม่ใช่ผู้ป่วย แต่เป็นผู้ให้การรักษา มีเจ้าหน้าที่เวชปฏิบัติ **103 คน** ถูกสุ่มให้เข้ากลุ่มใช้เครื่องมือ 52 คน และกลุ่มควบคุม 51 คน ทั้งสองกลุ่มใช้เวชระเบียนอิเล็กทรอนิกส์บนระบบคลาวด์แบบเดียวกัน แต่กลุ่มทดลองมีเครื่องมือ **AI Consult เวอร์ชัน 2.0** เพิ่มเข้ามา เครื่องมือนี้สร้างบนโมเดลภาษาขนาดใหญ่ **GPT-4o ของ OpenAI** และฝังอยู่ในระบบเวชระเบียน มันอ่านข้อความที่ผู้ให้การรักษาบันทึกไว้ แล้วแจ้งเตือนเมื่อพบประเด็นที่อาจผิดพลาดหรือควรพิจารณาเพิ่มเติมในส่วนของการวินิจฉัยหรือแผนการรักษา ผู้ให้การรักษายังคงมีอำนาจตัดสินใจเต็มที่ว่าจะรับ แก้ไข หรือเพิกเฉยต่อคำแนะนำของระบบ

> **ใช้โมเดลอะไร และทำไมรายละเอียดนี้จึงสำคัญ**
>
> เครื่องมือที่ใช้คือ **AI Consult 2.0** ซึ่งทำงานด้วย **GPT-4o ของ OpenAI** รุ่นเดือนพฤษภาคม 2025 ผ่าน API เชิงพาณิชย์ของ OpenAI ภายใต้สัญญาสำหรับองค์กร และตั้งค่าความสุ่มไว้ต่ำ (temperature 0.1) เครื่องมือถูกฝังในเวชระเบียนอิเล็กทรอนิกส์ที่พัฒนาขึ้นเฉพาะสำหรับ EasyClinic และใช้คำสั่งระบบที่เขียนให้สอดคล้องกับแนวทางการรักษาแห่งชาติของเคนยา ผู้เขียนเผยแพร่คำสั่งเต็มที่ใช้กับโมเดลด้วย
> เหตุใดรายละเอียดนี้จึงสำคัญ? เพราะผลการทดลองนี้พูดถึง **ระบบหนึ่งระบบที่กำหนดไว้ชัดเจน** — โมเดลรุ่นหนึ่ง คำสั่งหนึ่งชุด เวชระเบียนระบบหนึ่ง และการตั้งค่าหนึ่งแบบ — ไม่ใช่ "LLM ในการแพทย์" ทั้งหมด ผู้เขียนย้ำประเด็นเดียวกัน โดยเรียกผลลัพธ์นี้ว่า *เกณฑ์อ้างอิงตามช่วงเวลา (temporal benchmark)* มากกว่าจะเป็นค่าคงที่ของความสามารถ โมเดลรุ่นใหม่กว่า คำสั่งต่างออกไป หรือคลินิกที่ใช้ระบบดิจิทัลน้อยกว่านี้ อาจให้ผลที่ต่างกันได้
> ด้านความเป็นอิสระของงาน OpenAI ภายหลังให้การสนับสนุนแบบไม่เป็นเงินสด เช่น เครดิตประมวลผลบนคลาวด์ และคำแนะนำทางเทคนิคเกี่ยวกับ API แต่ผู้เขียนระบุว่าตัดสินใจเลือกใช้ OpenAI **ก่อน** ข้อเสนอสนับสนุนนั้น และ OpenAI ไม่มีบทบาทในการออกแบบการทดลอง เก็บข้อมูล วิเคราะห์ข้อมูล หรือการตัดสินใจตีพิมพ์

ระหว่างวันที่ 22 เมษายนถึง 16 กรกฎาคม 2025 มีผู้ป่วยเข้าร่วม **9,691 คน** ผลลัพธ์หลักถูกออกแบบให้ยึดผู้ป่วยเป็นศูนย์กลางและค่อนข้างเข้มงวด คือ **ผลรวมของเหตุการณ์ที่ถือว่าการรักษาล้มเหลวภายใน 14 วัน** หลังมารับบริการ โดยคณะผู้เชี่ยวชาญเป็นผู้ตัดสินผลลัพธ์แต่ละรายโดยไม่ทราบว่าผู้ป่วยอยู่ในกลุ่มใด เช่น อาการไม่ดีขึ้นหรือโรคแย่ลง การทดลองลงทะเบียนไว้ล่วงหน้าใน Pan-African Clinical Trials Registry หมายเลข 202502499779176

การเลือกผลลัพธ์แบบนี้สำคัญมาก เพราะการแสดงว่า AI ทำให้สิ่งที่ผู้ให้การรักษา **เขียน** เปลี่ยนไปนั้นทำได้ค่อนข้างง่าย แต่การแสดงว่า AI ทำให้สิ่งที่ **เกิดขึ้นกับผู้ป่วย** เปลี่ยนไปนั้นยากกว่า และมีความหมายมากกว่า

พบอะไร

ผลลัพธ์หลักไม่ได้ดีขึ้นอย่างมีนัยสำคัญ การรักษาล้มเหลวเกิดในผู้ป่วย 102 จาก 4,693 คน (2.2%) ในกลุ่ม AI และ 94 จาก 4,654 คน (2.0%) ในกลุ่มควบคุม ตัวเลขดิบสูงกว่าเล็กน้อยในกลุ่ม AI แต่เมื่อปรับความแตกต่างระหว่างกลุ่มผู้ให้การรักษาแล้ว ค่าประมาณกลับเอนไปในทิศที่อาจเป็นประโยชน์ โดยมี **อัตราส่วนออดส์ที่ปรับแล้ว 0.77 (ช่วงความเชื่อมั่น 95% 0.55–1.08, P = 0.13)** ซึ่งไม่ถึงระดับนัยสำคัญทางสถิติ การที่ตัวเลขดิบกับค่าที่ปรับแล้วชี้คนละทิศไม่ได้เกิดจากการคำนวณผิด แต่เกิดจากการปรับความแตกต่างระหว่างกลุ่มผู้ให้การรักษา ไม่ว่าจะอ่านแบบใด ช่วงความเชื่อมั่นยังครอบคลุมกรณี "ไม่มีผล" อย่างชัดเจน จึงยังอ้างประโยชน์ต่อผู้ป่วยไม่ได้ และขนาดผลโดยสัมบูรณ์ก็เล็กมาก

หากต้องการวิธีอ่านอัตราส่วนออดส์ ช่วงความเชื่อมั่น และค่า P ร่วมกันแบบภาษาปกติ ดูคู่มืออ่านผลการทดลองทางคลินิก

ไม่พบสัญญาณด้านความปลอดภัย — แต่มีขอบเขตของข้อสรุป ไม่มีเหตุการณ์ไม่พึงประสงค์ร้ายแรงใดที่ถูกตัดสินว่าเกี่ยวข้องกับเครื่องมือ และการทบทวนอิสระก็ไม่พบสัญญาณด้านความปลอดภัย อย่างไรก็ตาม ผู้เขียนระบุข้อจำกัดไว้อย่างตรง

ไปตรงมา การทดลองไม่ได้มีขนาดใหญ่พอที่จะตรวจหาอันตรายร้ายแรงที่เกิดขึ้นได้น้อย และไม่มีกรอบความไม่ด้อยกว่าหรือกรอบความปลอดภัยที่กำหนดไว้ล่วงหน้า ดังนั้นงานนี้จึง **ไม่สามารถพิสูจน์** ความปลอดภัยสำหรับเหตุการณ์หายากได้

การบันทึกข้อมูลทางคลินิกดีขึ้น จากการพบผู้ป่วย 2,000 ครั้งที่ผู้เชี่ยวชาญทบทวนโดยไม่ทราบกลุ่ม ผู้ให้การรักษาที่ใช้ AI Consult มีคุณภาพการบันทึกข้อมูลดีขึ้นในทุกด้านที่ให้คะแนน ทั้งการวินิจฉัยที่บันทึกไว้ แผนการรักษา และความครบถ้วนโดยรวม

การสั่งยาแทบไม่เปลี่ยน ไม่พบความแตกต่างอย่างมีนัยสำคัญในการสั่งยา รวมถึงความถูกต้องของการใช้ยาปฏิชีวนะ (อัตราส่วนออดส์ที่ปรับแล้ว 0.86, ช่วงความเชื่อมั่น 95% 0.48–1.55) เครื่องมือไม่ได้เปลี่ยนอัตราการสั่งยาปฏิชีวนะอย่างชัดเจน

ผู้ป่วยไม่รู้สึกว่าประสบการณ์ต่างกัน ในผู้ป่วย 826 คนที่ตอบแบบสำรวจ ความพึงพอใจแทบเหมือนกันระหว่างสองกลุ่ม และเวลาที่ใช้ในการปรึกษาก็ใกล้เคียงกัน

ต้นทุนบางส่วนมีแนวโน้มลดลง ในการวิเคราะห์ที่ปรับปัจจัยกวนแล้ว ค่าใช้จ่ายที่เกี่ยวข้องกับยาปฏิชีวนะต่ำกว่าในกลุ่ม AI ซึ่งอาจเกิดจากการเลือกยาที่ราคาถูกกว่า มากกว่าการสั่งยาน้อยลง เงินที่ประหยัดได้ต่อผู้ป่วยดูเหมือนจะมากกว่าค่าใช้เครื่องมือต่อคน แต่ผู้เขียนระบุชัดว่านี่ยังเป็นเพียงสัญญาณเบื้องต้น ไม่ใช่ข้อสรุปทางเศรษฐศาสตร์สุขภาพ เพราะการคำนวณต้นทุนรวมตลอดการใช้งานอยู่นอกขอบเขตของการทดลอง

สรุปของผู้เขียนเองตรงที่สุด: ภายในขอบเขตของการทดลอง เครื่องมือช่วยด้วย LLM ดูปลอดภัย แต่ไม่ได้ลดการรักษาล้มเหลวภายใน 14 วัน และหากมีประโยชน์จริง ขนาดของประโยชน์ก็น่าจะไม่มาก

ทำไม “ไม่แตกต่างอย่างมีนัยสำคัญ” ไม่เท่ากับ “ใช้ไม่ได้”

ผลลัพธ์หลักที่ไม่แตกต่างอย่างมีนัยสำคัญสามารถถูกอ่านเกินจริงได้ทั้งสองทิศ มีสองเหตุผลสำคัญที่ทำให้เรื่องไม่ง่ายขนาดนั้น

ประการแรก **การทดลองนี้ออกแบบมาเพื่อจับผลที่ใหญ่กว่าผลที่พบจริง** เหตุการณ์ร้ายแรงในงานปฐมภูมิเกิดไม่บ่อย — ที่นี่ประมาณ 2% — ดังนั้นการแยกประโยชน์เล็ก ๆ ที่มีอยู่จริงออกจากความผันผวนแบบสุ่มต้องใช้ผู้เข้าร่วมจำนวนมากมาก การคำนวณกำลังทางสถิติภายหลังของผู้เขียนชี้ว่า หากต้องการตรวจพบผลขนาดเท่าที่เห็นในงานนี้ อาจต้องใช้การทดลองที่มีผู้ป่วย**มากกว่า 100,000 คน** ดังนั้นผลที่ไม่ถึงนัยสำคัญในงานขนาดนี้ไม่ได้ตัดความเป็นไปได้ของประโยชน์เล็ก ๆ ออก เพียงแต่การทดลองนี้ยังไม่มีความละเอียดพอที่จะยืนยันได้

ประการที่สอง **ความแตกต่างระหว่างสองกลุ่มถูกทำให้พร่าลงบางส่วน** ความผิดพลาดในการตั้งค่าระบบช่วงสั้น ๆ ทำให้ผู้ให้การรักษาบางคนในกลุ่มควบคุมเข้าถึง AI Consult ได้ นอกจากนี้ ผู้ให้การรักษาในเครือข่ายเดียวกันย่อมพูดคุยและอาจนำวิธีปฏิบัติใหม่ข้ามกลุ่มไปด้วย ปัจจัยทั้งสองทำให้สองกลุ่มดูคล้ายกันมากขึ้น และผลักความแตกต่างจริงให้เข้าใกล้ศูนย์ ยิ่งไปกว่านั้น เครือข่ายที่ทำการทดลองมีมาตรฐานการดูแลค่อนข้างสูงอยู่แล้ว จึงเหลือพื้นที่ให้เครื่องมือแสดงการปรับปรุงได้น้อย

สิ่งเหล่านี้ไม่ได้ช่วยให้ตั้งพาดหัวว่า “พลิกวงการ” แต่หมายความว่าควรอ่านผลอย่างได้สัดส่วน บนเกณฑ์ผลลัพธ์ที่ยากและตรงที่สุด เครื่องมือนี้ **ยังไม่ได้แสดงว่าช่วยผู้ป่วยได้ภายในสองสัปดาห์** ขณะเดียวกันก็ช่วยให้การบันทึกข้อมูลดีขึ้นอย่างวัดได้ และไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจน

สิ่งที่ผลนี้ *ยังพิสูจน์ไม่ได้*

- **ไม่ได้พิสูจน์ว่า AI ช่วยให้ผลลัพธ์ของผู้ป่วยดีขึ้น** บนผลลัพธ์หลักที่ 14 วัน ไม่พบประโยชน์อย่างมีนัยสำคัญ
- **ไม่ได้พิสูจน์ว่า AI ไม่มีประโยชน์** ค่าประมาณเอนไปในทิศที่อาจเป็นประโยชน์ การบันทึกข้อมูลดีขึ้นทุกด้าน และค่ายามี

แนวโน้มลดลง ผลที่ไม่แตกต่างอย่างมีนัยสำคัญยังสอดคล้องกับประโยชน์เล็ก ๆ ที่การทดลองนี้มีขนาดไม่พอจะยืนยัน

- **ไม่ได้พิสูจน์ว่าเครื่องมือปลอดภัยต่ออันตรายที่พบได้น้อย** แม้ไม่พบสัญญาณด้านความปลอดภัย แต่การทดลองไม่ได้ออกแบบให้ตรวจหาเหตุการณ์ร้ายแรงที่เกิดไม่บ่อย
- **ไม่ได้แสดงว่า "AI ชนะหมอ" หรือแทนผู้ให้การรักษาได้** นี่เป็นเครื่องมือช่วยตัดสินใจ ผู้ให้การรักษายังคงมีอำนาจเต็มที่จะรับหรือปฏิเสธคำแนะนำ
- **นำผลไปใช้กับทุกบริบทโดยอัตโนมัติไม่ได้** การทดลองทำในเครือข่ายเอกชนในเขตเมืองแห่งเดียวในเคนยา ผลในพื้นที่ชนบท ชานเมือง หรือประเทศรายได้สูงอาจต่างออกไป
- **ยังยืนยันการประหยัดต้นทุนไม่ได้** สัญญาณด้านต้นทุนยังเป็นเพียงข้อมูลประกอบ ไม่ใช่การประเมินเศรษฐศาสตร์สุขภาพที่สมบูรณ์

หลักฐานน่าเชื่อถือเพียงใด

สำหรับข้อสรุปหลักว่า **ยังไม่มีหลักฐานว่าลดอันตรายระยะสั้นต่อผู้ป่วย** การออกแบบงานถือว่าแข็งแรง และข้อสรุปก็ระมัดระวังเหมาะสม การทดลองแบบไปข้างหน้า ลงทะเบียนล่วงหน้า สุ่มเป็นกลุ่ม และใช้ผลลัพธ์ระดับผู้ป่วยที่คณะผู้เชี่ยวชาญตัดสินโดยไม่ทราบกลุ่ม ถือว่าใกล้เคียงกับหลักฐานโลกจริงที่ดีที่สุดชนิดหนึ่งสำหรับเครื่องมือแบบนี้ และให้ข้อมูลมากกว่าคะแนนจากชุดทดสอบมาตรฐานหรือกรณีตัวอย่างจำลองอย่างมาก

สำหรับผลลัพธ์รอง — การบันทึกข้อมูลดีขึ้น การสั่งยาไม่เปลี่ยน ความพึงพอใจใกล้เคียง และค่ายาปฏิชีวนะลดลง — หลักฐานค่อนข้างดี แต่ควรอ่านในฐานะผลรอง ไม่ใช่พาดหัวหลัก และยังได้รับผลจากข้อจำกัดเรื่องการปนกันระหว่างกลุ่มและการทำงานในเครือข่ายเดียว

ด้านความปลอดภัย หลักฐานให้ความสบายใจได้ระดับหนึ่ง แต่มีขอบเขต ไม่พบสัญญาณอันตราย ทว่างานไม่ได้ออกแบบมาเพื่อค้นหาอันตรายที่พบได้น้อย

ท่าทีที่เป็นประโยชน์ที่สุดจึงไม่ใช่ "มันใช้ได้" หรือ "มันล้มเหลว" แต่คือ: การทดลองโลกจริงที่ออกแบบอย่างระมัดระวังพบว่าเครื่องมือ AI นี้ไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจนและช่วยกระบวนการดูแลบางส่วน แต่ยังไม่แสดงประโยชน์ต่อผลลัพธ์ของผู้ป่วยภายในสองสัปดาห์ และหากต้องการตรวจหาประโยชน์ขนาดเล็กเช่นนั้น อาจต้องใช้การทดลองที่ใหญ่กว่านี้มาก

ทำไมจึงสำคัญ

การถกเถียงเรื่อง AI ทางการแพทย์ขาดหลักฐานแบบนี้อย่างมาก มีงานวิจัยนับพันชิ้นที่แสดงว่าโมเดลทำข้อสอบได้ดีหรือทำผลงานใกล้เคียงผู้เชี่ยวชาญในกรณีตัวอย่างที่สะอาด แต่มีการทดลองแบบสุ่มขนาดใหญ่ในโลกจริงน้อยมากที่ถามว่าผู้ป่วยจริงดีขึ้นหรือไม่ งานนี้เป็นหนึ่งในนั้น และผลลัพธ์พากลับมาสู่ความจริงที่ไม่หวือหวา: **ทำข้อสอบผ่านไม่เท่ากับช่วยผู้ป่วยได้**

ช่องว่างนี้คือประเด็นสำคัญ เครื่องมือหนึ่งอาจมีประโยชน์จริงต่อผู้ให้การรักษา — บันทึกชัดขึ้น ค่ายาต่ำลง มีตัวช่วยตรวจทานอีกชั้น — แต่ยังไม่เปลี่ยนผลลัพธ์สำคัญของผู้ป่วยในสองสัปดาห์ ทั้งสองอย่างสามารถเป็นจริงพร้อมกันได้ ระบบสุขภาพที่มีวุฒิภาวะต้องรับความจริงทั้งสองด้าน ไม่ใช่เลือกด้านที่สะดวกกว่า

งานนี้ยังย้ายภาระการพิสูจน์ไปในทิศที่มีประโยชน์ หากบริษัทต้องการอ้างว่า AI ทางคลินิกช่วยปรับปรุงการดูแล หลักฐานที่เกี่ยวข้องไม่ใช่ตารางอันดับของโมเดล แต่คือการทดลองแบบนี้ บนผลลัพธ์ที่สำคัญต่อผู้ป่วย — และถ้าเป็นไปได้ ควรใหญ่กว่านี้ เพราะบทเรียนตรง ๆ ของงานนี้คือ ประโยชน์ที่มีขนาดไม่มากต้องใช้การศึกษาขนาดใหญ่มากจึงจะมองเห็น

สรุป

การทดลองเชิงปฏิบัติแบบสุ่มเป็นกลุ่มในสถานพยาบาลปฐมภูมิ 16 แห่งของเคนยาทดสอบเครื่องมือ AI เชิงกำเนิดสำหรับช่วยตัดสินใจชื่อ **AI Consult** ซึ่งเพิ่มเข้าไปในเวชระเบียนอิเล็กทรอนิกส์ที่เจ้าหน้าที่เวชปฏิบัติใช้อยู่ ในผู้ป่วย 9,691 คน ผลรวมของเหตุการณ์ที่ถือว่าการรักษาล้มเหลวภายใน 14 วัน ซึ่งผู้เชี่ยวชาญเป็นผู้ตัดสินโดยไม่ทราบกลุ่ม เกิด 2.2% เมื่อใช้เครื่องมือ เทียบกับ 2.0% เมื่อไม่ใช้ (อัตราส่วนออดส์ที่ปรับแล้ว 0.77, ช่วงความเชื่อมั่น 95% 0.55–1.08, P = 0.13) ซึ่งไม่แตกต่างอย่างมีนัยสำคัญ เครื่องมือไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจน ช่วยให้การบันทึกข้อมูลทางคลินิกดีขึ้นในทุกด้านที่ประเมิน ไม่เปลี่ยนการสั่งยา ไม่เปลี่ยนความพึงพอใจของผู้ป่วย และสัมพันธ์กับค่าใช้จ่ายด้านยาปฏิชีวนะที่ลดลงเล็กน้อย ผู้เขียนสรุปว่าเครื่องมือดูปลอดภัยภายในขอบเขตของงานนี้ แต่ไม่ได้ลดการรักษาล้มเหลว และหากมีประโยชน์จริงก็น่าจะมีขนาดไม่มาก การตรวจพบผลขนาดที่สังเกตได้อาจต้องใช้การทดลองระดับประมาณ 100,000 คน งานนี้จึงไม่ได้แสดงว่า AI ทางคลินิกช่วยให้ผลลัพธ์ของผู้ป่วยดีขึ้น และก็ไม่ได้แสดงว่าไม่มีประโยชน์เลย — มันแสดงว่าเครื่องมือที่ดูปลอดภัยและช่วยกระบวนการทำงานบางส่วน ยังไม่ได้ช่วยผู้ป่วยอย่างพิสูจน์ได้ภายในสองสัปดาห์ในเครือข่ายเอกชนเขตเมืองแห่งหนึ่ง

ประเมินแบบไม่เกินจริง

งานวิจัยแสดงอะไร: ในการทดลองแบบสุ่มในโลกจริง การเพิ่มเครื่องมือช่วยตัดสินใจที่ใช้ LLM เข้าไปในเวชระเบียนปฐมภูมิไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจน ช่วยให้คุณภาพการบันทึกข้อมูลทางคลินิกดีขึ้น ไม่เปลี่ยนการสั่งยา และไม่ลดผลรวมของเหตุการณ์การรักษาล้มเหลวภายใน 14 วันอย่างมีนัยสำคัญ

อะไรที่เป็นไปได้แต่ยังไม่พิสูจน์: เครื่องมืออาจลดการรักษาล้มเหลวได้จริงเพียงเล็กน้อย แต่การทดลองนี้มีขนาดไม่พอจะตรวจพบ อาจประหยัดเงินเมื่อรวมต้นทุนทั้งหมด และการบันทึกข้อมูลที่ดีขึ้นอาจนำไปสู่การดูแลที่ดีขึ้นในระยะยาว

มันไม่ได้แสดงอะไร: ไม่ได้แสดงว่า AI ทางคลินิกช่วยให้ผลลัพธ์ของผู้ป่วยดีขึ้น ไม่ได้พิสูจน์ความปลอดภัยต่อเหตุการณ์ร้ายแรงที่พบได้น้อย ไม่ได้แสดงว่า AI แทนหรือทำผลงานเหนือผู้ให้การรักษาได้ และยังไม่ยืนยันว่าผลนี้ใช้ได้กับพื้นที่ชนบทหรือประเทศรายได้สูง รวมทั้งยังไม่ยืนยันการประหยัดต้นทุน

ข้อจำกัดหลัก: การทดลองถูกออกแบบให้ตรวจจับผลที่ใหญ่กว่าที่พบจริง เหตุการณ์หลักเกิดไม่บ่อยจึงต้องใช้ตัวอย่างขนาดใหญ่มาก ความผิดพลาดในการตั้งค่าทำให้กลุ่มควบคุมบางส่วนเข้าถึงเครื่องมือ และการทำงานในเครือข่ายเดียวกันอาจทำให้พฤติกรรมข้ามระหว่างกลุ่ม ทั้งสองปัจจัยทำให้ความแตกต่างถูกดึงเข้าหาศูนย์ งานทำในเครือข่ายเอกชนในเขตเมืองแห่งเดียวที่มีมาตรฐานค่อนข้างสูง ไม่มีกรอบความไม่ด้อยกว่าหรือกรอบความปลอดภัยที่กำหนดไว้ล่วงหน้า และติดตามผลเพียง 14 วัน

ผู้อ่านทั่วไปควรมั่นใจแค่ไหน? มั่นใจได้ค่อนข้างสูงว่าในการทดลองนี้เครื่องมือไม่ก่อให้เกิดสัญญาณด้านความปลอดภัยที่ชัดเจนและช่วยให้การบันทึกข้อมูลดีขึ้น มั่นใจได้สูงเช่นกันว่างานนี้ยังไม่ได้แสดงว่าผลลัพธ์ของผู้ป่วยภายใน 14 วันดีขึ้น ส่วนข้ออ้างกว้าง ๆ ว่าเครื่องมือนี้ “ได้ผล” หรือ “ล้มเหลว” ในการช่วยผู้ป่วยยังมีความมั่นใจต่ำ คำถามนั้นยังเปิดอยู่และต้องการการทดลองที่ใหญ่กว่ามาก

แหล่งข้อมูล

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