The CLEAN PAPER

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

AI קליני שנראה בטוח ושיפר את התיעוד — אבל לא שיפר תוצאות מטופלים

ניסוי פרגמטי, אקראי באשכולות, ב-*16* מתקני טיפול ראשוני בקניה בדק כלי תמיכה בהחלטות מבוסס *AI* גנרטיבי שנוסף לרשומה שבה משתמשים *officers. clinical* בקרב *9,691* מטופלים, ה-*composite* המחמיר ל-*14* יום של כישלון טיפול שנשפט בידי מומחים היה *2.2%* עם הכלי לעומת *2.0%* בלעדיו *odds (adjusted* *0.77 ratio, 95% CI 0.55–1.08, P = 0.13)* — ללא הבדל מובהק. הכלי לא הראה אות בטיחות, שיפר תיעוד בכל התחומים, הותיר מרשמים ושביעות רצון ללא שינוי ונטה לעלויות אנטיביוטיקה נמוכות יותר. זה אינו מחקר שנכשל; זה מחקר נדיר וכנה — שנמדד על מטופלים, לא על *benchmarks* — והוא מגיע לאמת הלא-זוהרת שכלי שנראה בטוח ועזר לתהליך לא עזר למטופלים באופן שניתן להדגים בתוך שבועיים, ושזיהוי תועלת כזאת ידרוש ניסוי גדול בהרבה.

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

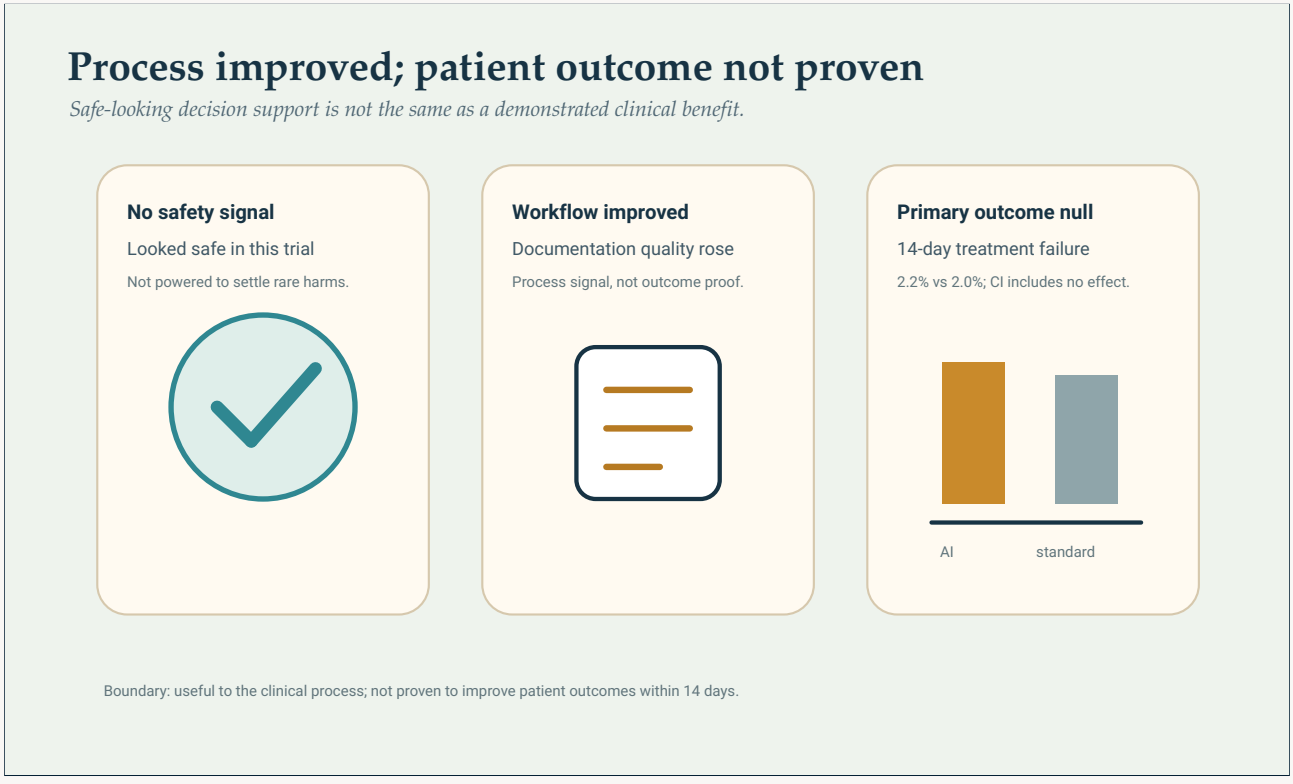
בטוח ומועיל זה לא אותו דבר כמו יעיל

רוב הכותרות על AI רפואי נבנות מסוג המחקר הלא נכון. מודל מקבל ציון גבוה בשאלות בחינה או מנצח רופאים בווינייטות שנבחרו בקפידה, והסיפור כותב את עצמו: המכונה מוכנה למרפאה.

הניסוי הזה עשה משהו קשה ונדיר יותר. הוא הכניס כלי תמיכה בהחלטות מבוסס AI גנרטיבי לטיפול ראשוני אמיתי, במתקנים אמיתיים, עם מטופלים אמיתיים, ושאל את השאלה שבאמת חשובה: האם מצב המטופלים השתפר?

התשובה הכנה היא לא — לא באופן שניתן למדוד, לא בתוך 14 יום. הכלי לא הראה אות בטיחות. הוא שיפר את איכות התיעוד הקליני. ייתכן שאפילו הפחית חלק מעלויות התרופות. אבל הוא לא הפחית באופן מובהק כישלונות טיפול, והמחברים מקפידים לומר שכל תועלת, אם קיימת, כנראה צנועה.

זו אינה כישלון של המחקר. זו עבודת המחקר. כך נראות ראיות אחראיות על AI ברפואה כאשר מודדים אותן על מטופלים במקום על benchmarks.



הכלי לא הראה אות בטיחות ושיפר את התיעוד הקליני, אבל תוצאת המטופל שנקבעה מראש ל-14 יום לא השתפרה באופן מובהק. עזרה לתהליך אינה זהה לתועלת מוכחת למטופל.

מה עשו המחברים

הצוות ערך ניסוי פרגמטי אקראי באשכולות ב-16 **מתקני טיפול ראשוני** שמפעילה רשת בריאות פרטית (Penda Health) במחוזות ניירובי וקיאמבו בקניה. הטיפול במתקנים האלה ניתן ברובו בידי **officers clinical** — אנשי מקצוע

בדרג ביניים בעלי דיפלומה של שלוש שנים — לעיתים בלי גישה קלה לייעוץ בכיר.

יחידת ההקצאה האקראית הייתה הקלינאי, לא המטופל. 103 officers clinical הוקצו באקראי: 52 לזרוע ההתערבות ו-51 לזרוע הביקורת. שתי הזרועות השתמשו באותה רשומה רפואית אלקטרונית בענן. זרוע ההתערבות קיבלה בנוסף את **Consult AI** (גרסה 2.0), כלי תמיכה בהחלטות שנבנה על מודל השפה הגדול **GPT-4o של OpenAI** והוטמע ברשומה. הוא קרא את המידע שהקלינאי תיעד ויכול היה לסמן בעיות אפשריות באבחנה או בתכנית הטיפול. הקלינאים שמרו על אוטונומיה מלאה: הם יכלו לקבל, לשנות או להתעלם מההצעות.

> **איזה מודל זה היה, ולמה הפרטים חשובים**
>
> היה הכלי **AI Consult 2.0**, מאי (מהדורת **OpenAI של GPT-4o** את שהריץ, של המסחרי ה-API דרך 2025), אלקטרונית רשומה בתוך ישב הוא 0.1). temperature) נמוכות אקראיות ובגדרות enterprise ברישיון OpenAI הטיפול להנחיות להתאים כדי שנכתבו system prompts באמצעות והונחה (EasyClinic של EMR) ייעודית המלא. ההוראות prompt את פרסמו המחברים קניה; של הלאומיות
> למה לפרט? משום שהתוצאה היא על מערכת מסוימת אחת — גרסת מודל אחת, prompt אחד, רשומה אחת, הגדרה אחת — ולא על "LLMs ברפואה" באופן כללי. המחברים מדגישים אותה נקודה: הם מכנים את הממצא נקודת ייחוס בזמן ולא אומדן קבוע של היכולת. מודל חדש יותר, prompt אחר או מרפאה פחות ממוחשבת יכולים כולם לשנות את התוצאה.
> לגבי עצמאות: OpenAI סיפקה מאוחר יותר תמיכה בעין (קרדיטי מחשוב ענן והכוונה טכנית לשימוש ב-API), אבל המחברים מציינים שההחלטה להשתמש ב-OpenAI התקבלה לפני ההצעה הזאת, וש-OpenAI לא הייתה מעורבת בתכנון הניסוי, באיסוף הנתונים, בניתוח או בהחלטה לפרסם.

בין 22 באפריל ל-16 ביולי 2025 נכללו **9,691 מטופלים**. **התוצאה הראשית הייתה בכוונה ממוקדת-מטופל ומחמירה**: composite של כישלון טיפול בתוך 14 יום, שנקבע בידי מומחים — פאנל קלינאים, בסמיות לזרוע המחקר, שפט אם לכל מטופל הייתה תוצאה רעה כמו מחלה שלא נפתרה או החמירה. הניסוי נרשם מראש (Pan-African Clinical Trials Registry 202502499779176).

בחירת התכנון הזאת היא כל העניין. קל להראות שכלי AI משנה את מה שקלינאי כותב. הרבה יותר קשה — והרבה יותר משמעותי — להראות שהוא משנה את מה שקורה למטופל.

מה הם מצאו

התוצאה הראשית לא השתפרה. כישלון טיפול התרחש אצל 102 מתוך 4,693 מטופלים (2.2%) בזרוע ה-AI ואצל 94 מתוך 4,654 (2.0%) בזרוע הביקורת. האחוזים הגולמיים היו מעט גבוהים יותר עם AI, אבל לאחר התאמה אומדן הנקודה נטה לכיוון תועלת: **יחס הסיכויים המתוקנן היה 0.77 (רווח ביטחון 95% 0.55 עד 1.08, P = 0.13)** — לא מובהק סטטיסטית. ההיפוך הזה בין המספרים הגולמיים למתוקננים אינו טעות חשבון; ההתאמה מתקנת להבדלים בין אשכולות הקלינאים. כך או כך, רווח הביטחון כולל בנוחות "אין אפקט", ולכן אי-אפשר לטעון לתועלת, ובמונחים מוחלטים האפקט היה זעיר.

להסבר פשוט על קריאת ratios, odds רווחי ביטחון וערכי P יחד, ראו המדריך לתוצאות קליניות.

לא נמצא אות בטיחות — בתוך גבולות. אף אירוע לוואי חמור לא נחשב קשור לכלי, ובדיקה עצמאית לא מצאה אות בטיחות. המחברים כנים לגבי גבול ההרגעה הזאת: הניסוי לא היה בעל כוח סטטיסטי לזהות נזקים חמורים נדירים, ולא הייתה לו מסגרת פורמלית לאי־נחיתות (noninferiority) או לבטיחות שנקבעה מראש, ולכן הוא אינו יכול להוכיח

בטיחות לאירועים לא שכיחים.

התיעוד השתפר. מתוך 2,000 מפגשים שנבדקו בידי מומחים בסמיות, קלינאים שהשתמשו ב-Consult AI הפיקו תיעוד קליני טוב יותר בכל התחומים שדורגו — האבחנה שנרשמה, תכנית הטיפול והשלמות הכללית.

המרשמים כמעט לא השתנו. לא היה הבדל מובהק במתן מרשמים, כולל שימוש נכון באנטיביוטיקה (adjusted odds ratio 0.86, 95% CI 0.48 עד 1.55). הכלי לא שינה את שיעור מתן האנטיביוטיקה.

המטופלים לא הרגישו הבדל. בקרב 826 מטופלים שהשלימו סקר שביעות רצון, שביעות הרצון הייתה כמעט זהה בשתי הזרועות, וזמני הייעוץ היו דומים.

העלויות נטו מעט מטה. בניתוח מתוקנן, עלויות הקשורות לאנטיביוטיקה היו נמוכות יותר בזרוע ה-AI — כנראה דרך בחירת תרופות זולות יותר ולא דרך פחות מרשמים — והחיסכון באנטיביוטיקה למטופל נראה גדול מעלות הפעלת הכלי למטופל. המחברים מסמנים זאת כרמז, לא כמסקנה: חשבונאות מלאה של total cost of ownership הייתה מחוץ לניסוי.

השורה האחת של המחברים עצמם היא הניסוח המדויק ביותר: עזרת LLM הייתה בטוחה בתוך הגבולות האלה אך לא הפחיתה כישלון טיפול בתוך 14 יום, וכל תועלת כנראה צנועה.

למה "אין הבדל מובהק" לא אומר "זה לא עובד"

קל לקרוא יתר על המידה תוצאה ראשית אפסית לשני הכיוונים. שני דברים מונעים את הסיפור הפשוט.

ראשית, **הניסוי נבנה לגלות אפקט גדול יותר מזה שנמצא.** תוצאות רעות חמורות בטיפול ראשוני נדירות — סביב 2% כאן — ולכן כדי להבחין בין תועלת אמיתית קטנה לבין רעש צריך מספרים עצומים. חישובי הכוח post-hoc של המחברים מצביעים על כך שזיהוי אפקט בגודל שנצפה ידרוש **ניסוי גדול בהרבה, בסדר גודל של יותר מ-100,000 מטופלים.** תוצאה לא-מובהקת במחקר בגודל הזה אינה שוללת תועלת אמיתית קטנה; פירושה שהמחקר הזה לא הצליח לפתור אותה.

שנית, **ההשוואה הייתה מטושטשת בחלקה.** שגיאת תצורה נתנה לזמן קצר לחלק מן הקלינאים בזרוע הביקורת גישה ל-Consult AI, וקלינאים באותה רשת מדברים זה עם זה ונושאים הרגלים מעבר לגבול בין הקבוצות. שני הדברים נוטים להפוך את הזרועות לדומות יותר, ולדחוף כל הבדל אמיתי לכיוון אפס. נוסף על כך, הרשת המארחת כבר פעלה בסטנדרטים גבוהים יחסית, כך שנותר פחות מקום לכלי להראות שיפור.

אף אחת מן הנקודות האלה אינה מצילה כותרת "פריצת דרך". אבל הן אומרות שהקריאה הנכונה היא מכוילת, לא מבטלת: על נקודת הסיום הקשה והכנה ביותר, הכלי לא עזר למטופלים באופן שניתן להדגים בתוך שבועיים — ובו בזמן הוא כן עזר לתיעוד באופן מדיד ונראה בטוח.

מה זה לא מוכיח

- הוא **לא** מראה שה-AI שיפר תוצאות מטופלים. בנקודת הסיום הראשית ל-14 יום לא הייתה תועלת מובהקת.
- הוא **לא** מראה שה-AI חסר תועלת. אומדן הנקודה נטה לטובתו, התיעוד השתפר בכל התחומים ועלויות התרופות נטו לרדת; התוצאה האפסית עולה בקנה אחד עם תועלת אמיתית קטנה שהניסוי היה קטן מדי לאשר.
- הוא **לא** מוכיח שהכלי בטוח לנזקים נדירים. לא נמצא אות בטיחות, אבל הניסוי לא היה גדול או מתוכנן כדי לאשר

בטיחות לאירועים חמורים לא שכיחים.

- הוא **לא** מראה ש"AI מנצח רופאים" או מחליף קלינאים. זו תמיכה בהחלטות; הקלינאי שמר סמכות מלאה לקבל או לדחות אותה.
- הוא **לא** מכליל אוטומטית. הניסוי נערך ברשת פרטית עירונית אחת בקניה; מסגרות כפריות, פרבריות או במדינות עשירות עשויות להתנהג אחרת לשני הכיוונים.
- הוא **לא** מוכיח חיסכון בעלויות. אות העלות הוא רמז, לא הערכה כלכלית מלאה.

עד כמה הראיות חזקות?

לגבי הטענה המרכזית — אין ירידה מוכחת בנזק קצר-טווח למטופל — הראיות חזקות מבחינת התכנון, וצנועות כראוי במסקנה. ניסוי פרוספקטיבי, רשום מראש, אקראי באשכולות, עם composite ברמת המטופל שנשפט בסמיות, קרוב לראיות העולם-האמיתי הטובות ביותר שאפשר לאסוף לכלי כזה. הוא אינפורמטיבי בהרבה מציון benchmark או מחקר וינייטות.

לגבי הממצאים המשניים — תיעוד טוב יותר, מרשמים ללא שינוי, שביעות רצון דומה, עלויות אנטיביוטיקה נמוכות יותר — הראיות טובות, אך צריך לקרוא אותן כמשניות: אותות תומכים, לא הכותרת, ופגיעים לאותן מגבלות של contamination ורשת יחידה.

לגבי בטיחות, הראיות מרגיעות אך תחומות: לא נמצא אות, אך זה לא מחקר שנבנה למצוא נזקים נדירים.

הגישה השימושית ביותר אינה "זה עובד" וגם לא "זה נכשל". היא: ניסוי עולם-אמיתי קפדני מצא שכלי ה-AI הזה לא העלה אות בטיחות ושיפר את תהליך הטיפול, בלי להדגים תועלת בתוצאות מטופלים בתוך שבועיים — ושכדי לזהות תועלת כזאת, אם קיימת, יידרש מחקר גדול בהרבה.

למה זה חשוב

הדיון על AI רפואי רעב בדיוק לסוג הזה של ראיות. יש אלפי מאמרים שמראים מודלים מצטיינים בבחינות ומשתווים לקלינאים במקרים מסודרים. יש מעט מאוד ניסויים גדולים, פרגמטיים ואקראיים שמודדים אם מטופלים אמיתיים יוצאים נשכרים. זה אחד מהם, והוא נוחת על האמת הלא-זוהרת: לעבור את המבחן זה לא אותו דבר כמו לעזור למטופל.

הפער הזה הוא כל הסיפור. כלי יכול להיות באמת שימושי לקלינאים — רשומות ברורות יותר, חשבונות תרופות נמוכים יותר, זוג עיניים נוסף — ועדיין לא להזיז תוצאת מטופל קשה בתוך שבועיים. שתי העובדות יכולות להיות נכונות בו-זמנית, ומערכת בריאות בוגרת צריכה להחזיק את שתיהן במקום לבחור את הנוחה.

הוא גם מאפס את נטל ההוכחה לכיוון שימושי. אם חברה רוצה לטעון שה-AI הקליני שלה משפר טיפול, הראיה הרלוונטית אינה leaderboard. היא ניסוי כזה, על תוצאות שחשובות למטופלים — ועדיף ניסוי גדול יותר, משום שהלקח הכנה כאן הוא שתועלות צנועות דורשות מחקרים גדולים כדי לראות אותן.

בקצרה

ניסוי פרגמטי, אקראי באשכולות, ב-16 מתקני טיפול ראשוני בקניה בדק כלי תמיכה בהחלטות מבוסס AI גנרטיבי ("AI Consult") שנוסף לרשומה האלקטרונית שבה משתמשים clinical officers. בקרב 9,691 מטופלים, ה-composite של כישלון טיפול בתוך 14 יום שנשפט בידי מומחים היה 2.2% עם הכלי לעומת 2.0% בלעדיו (adjusted odds ratio 0.77, 95% CI 0.55–1.08, P = 0.13) — ללא הבדל מובהק. הכלי לא הראה אות בטיחות, שיפר תיעוד קליני בכל התחומים שדורגו, לא שינה מרשמים, הותיר שביעות רצון ללא שינוי והיה קשור לעלויות מעט נמוכות יותר של אנטיביוטיקה. המחברים מסיקים שבתוך הגבולות האלה הוא היה בטוח אך לא הפחית כישלון טיפול, וכל תועלת כנראה צנועה; זיהוי אפקט בגודל שנצפה ידרוש ניסוי גדול בהרבה (בסדר גודל של 100,000 מטופלים). התוצאה לא מראה ש-AI קליני משפר תוצאות מטופלים, וגם לא שהוא חסר תועלת — היא מראה שכלי שנראה בטוח ועזר לתהליך לא עזר למטופלים באופן שניתן להדגים בתוך שבועיים, ברשת פרטית עירונית אחת.

בדיקה מפוכחת

מה המאמר מראה: בניסוי אקראי בעולם האמיתי, הוספת כלי תמיכה בהחלטות מבוסס LLM לרשומות בטיפול ראשוני לא העלתה אות בטיחות, שיפרה את איכות התיעוד הקליני, לא שינתה מרשמים ולא הפחיתה באופן מובהק composite מחמיר של כישלון טיפול בתוך 14 יום.

מה סביר אך לא הוכח: שהכלי מפחית במידה קטנה את כישלון הטיפול, אפקט שהיה קטן מדי לניסוי הזה; שהוא חוסך כסף כאשר סופרים את כל העלויות; שתיעוד טוב יותר מתורגם בסופו של דבר לטיפול טוב יותר.

מה הוא לא מראה: ש-AI קליני משפר תוצאות מטופלים; שהוא בטוח לאירועים נדירים; שהוא מחליף קלינאים או עולה עליהם; שהתוצאות עוברות למסגרות כפריות או עשירות יותר; או שהחיסכון בעלויות הוכח.

המגבלות העיקריות: המחקר היה בעל כוח לאפקט גדול יותר מזה שנצפה (תוצאות נדירות דורשות מדגמים גדולים מאוד); שגיאת תצורה זיהמה את זרוע הביקורת והרגלים בתוך הרשת מטשטשים את ההשוואה, שניהם מטות לכיוון אין הבדל; רשת פרטית עירונית אחת שכבר פועלת בסטנדרטים גבוהים; אין מסגרת noninferiority או בטיחות שנקבעה מראש; אופק קצר של 14 יום.

כמה ביטחון צריך להיות לקורא כללי? גבוה בכך שהכלי לא הראה אות בטיחות בניסוי הזה ושיפר תיעוד. גבוה בכך שהוא לא שיפר באופן שניתן להדגים תוצאות מטופלים ל-14 יום כאן. נמוך לכל טענה שהוא "עובד" או "נכשל" כהתערבות לתועלת המטופל — השאלה הזאת באמת לא פתורה ודורשת מחקר גדול בהרבה. הגישה המתאימה: תוצאה זהירה בעולם האמיתי על כלי שלא העלה אות בטיחות ושיפר את תהליך הטיפול, לא פסק דין ש-AI משנה — או הורס — את הטיפול הראשוני.

מקורות

(2026) Medicine Nature al., et Agweyu
https://doi.org/10.1038/s41591-026-04503-6

202502499779176 registration Registry, Trials Clinical Pan-African
https://pactr.samrc.ac.za/

(2026) trial the on release press PATH / EurekAlert
https://www.eurekalert.org/news-releases/1133583