TELEHEALTH AND MEDICINE TODAY

ORIGINAL RESEARCH

International Telemedicine Program: Physician Versus AI Responses

Paul Hart, MD

Formerly Volunteer Physician, University of Massachusetts Medical School, Worcester, Massachusetts, USA

DOI: https://doi.org/10.30953/thmt.v11.712

Corresponding Author: Paul Hart, Email: paulhart46@yahoo.com

Keywords: artificial intelligence, asynchronous consultations, clinical decision support, diagnostic reasoning, international health, low-resource settings, OpenEvidence, patient safety, physician–AI comparison, telemedicine

Abstract

Background: Asynchronous, provider-to-provider telemedicine is increasingly used to extend specialist expertise to frontline clinicians in low-resource settings, yet the role of artificial intelligence (AI) clinical decision support in these workflows remains unclear. The Telemedicine Health and Empowerment Program of a U.S.-based non-governmental organization links the U.S.-based volunteer physicians with African providers through a secure, store-and-forward platform. OpenEvidence is an AI decision support tool that generates differentials and management suggestions from structured clinical narratives. This study compared how OpenEvidence and human volunteers responded to the same real-world telemedicine consultations.

Method: A retrospective comparative case series of 50 telemedicine consultations from African providers that had de-identified clinical narratives and written responses from volunteer physicians was conducted. The same narratives were entered into OpenEvidence, and AI outputs were captured without additional prompting. For each physician and AI, the response was recorded for the presence of four elements: (1) explicit diagnostic list or working diagnosis, (2) recommended next diagnostic steps (history, examination, laboratory tests, or procedures), (3) immediate treatment recommendations, and (4) identification of red flag features or follow-up concerns. Elements were compared descriptively between AI and physician responses.

Results: Seven consultations (14%) asked whether an initial assessment was correct, 9 (18%) requested guidance on further workup, and 34 (68%) sought a broader differential diagnosis. AI responses included a diagnostic list in 41 cases (82%); suggested next diagnostic steps in 50 (100%); immediate treatment recommendations in 50 (100%), red flag features in 26 (52%), and follow-up concerns in 37 (74%). Physician responses included an explicit diagnostic list in 12 cases (24%), suggested next diagnostic steps in 42 (84%), immediate treatment recommendations in 29 (58%), red flag features in 4 (8%), and follow-up concerns in 9 (18%).

Conclusions: In this African–U.S. telemedicine program, AI-generated responses were more structurally complete and safety-oriented, whereas physicians offered more selective, context-focused guidance. These findings support a hybrid telehealth workflow in which AI provides systematic completeness and safety checks while human consultants supply contextual judgment and local prioritization.

Plain Language Summary

There were 50 telemedicine consultations from African providers, which resulted in the following recommendations:

- explicit diagnostic list or working diagnosis
- recommended next diagnostic steps (history, examination, laboratory tests, or procedures)
- immediate treatment recommendations
- identification of red flag features or follow-up concerns.

In this African–U.S. telemedicine program, artificial intelligence-generated responses were more structurally complete and safety-oriented, whereas physicians offered more selective, context-focused guidance. These findings support a hybrid telehealth workflow in which AI provides systematic completeness and safety checks while human consultants supply contextual judgment and local prioritization.

Submitted: April 10, 2026, Accepted: May 26, 2026, Published: June 25, 2026

Asynchronous, provider-to-provider telemedicine is increasingly used to extend the expertise of specialists as frontline clinicians in low-resource settings.[1–4] Yet the role of artificial intelligence (AI) clinical decision support in these workflows remains unclear.[5,6] The Telemedicine Health and Empowerment Program of a U.S.-based non-governmental organization (NGO) links U.S.-based volunteer physicians with African providers through a secure, store-and-forward platform.[3,7] OpenEvidence is an AI decision support tool that generates differentials and management suggestions from structured clinical narratives.[5] This study compares how OpenEvidence and human volunteers responded to the same real-world telemedicine consultations.

This study examines how a clinical decision support platform (OpenEvidence) and volunteer physicians respond to the same real-world telemedicine consultations originating from frontline clinicians in Africa within the Telemedicine Health and Empowerment Program of a U.S. NGO. The work directly addresses the evolving role of AI in global telehealth and clinical decision support in low-resource settings.

International telemedicine programs that link specialists in high-income countries with clinicians in resource-limited

environments have become an important strategy for expanding access to expert guidance where specialist physicians are scarce. Within the clinic's Telemedicine Health and Empowerment Program, more than 100 U.S.-based physician volunteers provide asynchronous written consultation to approximately 1,350 African clinicians, including clinical officers and other frontline providers. Against this backdrop, AI-supported clinical decision tools such as OpenEvidence present a potential adjunct, but there is little empirical evidence on how their outputs compare with those of human consultants when applied to actual cases from low-resource telemedicine programs.

In this retrospective comparative case series of 50 telemedicine consultations, each de-identified case narrative and its original physician response were paired with an AI-generated response created by entering the same narrative into OpenEvidence. Each provider's question was categorized as (1) confirming an initial assessment, (2) seeking guidance on additional information to gather, or (3) requesting a broader differential diagnosis. For both AI and physician responses for the presence of four key elements: an explicit diagnostic list or working diagnosis; recommended next steps in data collection (history, examination, laboratory tests, or procedures); immediate treatment or initial management recommendations; and identification of red-flag features or follow-up concerns.

The AI responses demonstrated a highly structured and comprehensive pattern: a diagnostic list appeared in 41 of 50 cases (82%), recommendations for further workup and immediate treatment guidance were present in all 50 responses (100%), red-flag features were identified in 26 (52%), and management or follow-up concerns were documented in 37 (74%). Physician responses, by contrast, were more selective and focused. An explicit diagnostic list appeared in 12 cases (24%); next diagnostic steps were suggested in 42 (84%), immediate treatment recommendations in 29 (58%), red-flag features in 4 (8%), and explicit follow-up concerns in 9 (18%). These differences suggest a natural division of labor: AI is well suited to providing thorough, checklist-like coverage of differentials, safety flags, and follow-up considerations, whereas physicians contribute context-sensitive, pragmatic recommendations that reflect the realities of resource-constrained practice. The manuscript discusses how hybrid human–AI workflows could combine these complementary strengths in cross-border telehealth for low-resource settings.

Provider-to-provider telemedicine has become an important strategy for extending specialist expertise to frontline clinicians in low-resource settings, where specialty care is often scarce. Asynchronous, store-and-forward models are particularly well suited to such environments because they can function despite bandwidth limitations, geographic dispersion, and variable staffing. The Telemedicine Health and Empowerment Program of the American NGO is one such model, linking volunteer physicians based primarily in the United States with frontline providers in Africa via a secure platform. Within this program, clinical officers and other frontline clinicians submit de-identified clinical narratives and receive written specialist guidance intended to refine diagnosis, optimize management, and strengthen local decision-making.

In parallel, AI-assisted clinical decision support tools have emerged as potential adjuncts to telehealth workflows. OpenEvidence is an AI platform that analyzes structured clinical narratives to generate evidence-based outputs, including differential diagnoses, suggested diagnostic next steps, initial management recommendations, and identification of red flag features or follow-up needs. Early evaluations suggest that OpenEvidence can augment clinician decision-making by providing transparent, literature-linked recommendations in primary care and specialty contexts. However, most work on AI clinical decision support has been conducted in high-income settings, often using simulated vignettes or image-based tasks, with limited data from community-based telemedicine serving low- and middle-income countries.

The alignment between AI and physician-generated advice is especially important in cross-border telehealth that supports task shifting to non-physician clinicians. In these contexts, incomplete differentials or missed safety concerns may have an outsized impact, yet few empirical studies have examined how AI and human consultants differ in the structure, completeness, and safety orientation of their recommendations for real-world teleconsultations.

This study addresses that gap by comparing, for 50 telemedicine cases from the Telemedicine Health and Empowerment Program, the content of written responses from volunteer physicians with parallel outputs generated by OpenEvidence using the same de-identified clinical narratives. Primary outcomes included the presence of four key elements in each response—(1) explicit diagnostic list or working diagnosis, (2) recommended next diagnostic steps (history, examination, laboratory tests, or procedures), (3) immediate treatment recommendations, and (4) identification of red flag features or follow-up concerns—and the frequency with which each element appeared in AI versus physician responses.

Methods

Study Design and Setting

A retrospective comparative case series of paired physician and AI responses from the Telemedicine Health and Empowerment Program of an American-based NGO, which connects U.S.-based volunteer physicians with African providers via asynchronous, provider-to-provider telemedicine, was conducted. Consultations are submitted through a secure platform as free text clinical narratives with structured fields for patient demographics and presenting problems.

Case Selection and Classification

A convenience sample of 50 consultations was selected from a defined study period. Eligibility to participate included a case that had to include (1) a clearly documented clinical presentation with presenting complaints, pertinent history, examination findings, and any available investigations, and (2) a substantive written response from a physician volunteer. Cases were excluded if documentation was incomplete, the clinical picture was too ambiguous to categorize, or the physician response was missing.

Each case was classified according to the primary intent of the Kenyan provider's question:

1. confirmation of whether the provider's initial assessment or working diagnosis was correct

2. guidance on what additional information should be obtained before proceeding or
3. request for a broader differential diagnosis to consider other possible etiologies.

Generation of AI Responses

For each included case, the de-identified clinical narrative—presenting complaint, relevant history, examination findings, and available test results—was entered directly into OpenEvidence. OpenEvidence uses curated evidence sources and probabilistic reasoning to generate differential diagnoses, suggested diagnostic steps, and management recommendations based on the input narrative. Text was formatted for clarity and stripped of personal identifiers before entry. No additional prompting or instructions were used beyond the narrative itself. The primary AI output for each case was exported as is and stored alongside the corresponding physician response.

Data Extraction and Analysis

We reviewed the physician's and AI responses for each case using a shared structured framework. For every response, we recorded the presence or absence of four elements:

1. an explicit list of possible diagnoses or a clearly stated working diagnosis
2. recommended next steps in data gathering (additional history questions, targeted examination maneuvers, laboratory tests, or procedures)
3. suggestions for immediate treatment or initial management
4. identification of red flag features or safety concerns, including indications for urgent referral, escalation of care, or close monitoring.

Data were summarized descriptively as counts and percentages for AI and physician responses. The observed differences in the light of existing literature on telehealth in low-resource settings and emerging evidence on AI clinical decision support were interpreted.

Ethics

All cases were de-identified before analysis, with the removal of names, exact dates, and other direct identifiers. The review was conducted as a secondary analysis of routine program data; institutional review board determination (e.g., exempt or nonhuman subjects research), according to your approval.

Results

Reasons for Telemedicine Requests

The 50 consultations reflected three distinct types of questions from Kenyan providers. Seven cases (14%) requested confirmation that the provider's initial assessment was correct. Nine cases (18%) sought guidance on what additional information to collect before making management decisions. The largest group—34 cases (68%)—of Kenyan providers requested a broader differential diagnosis in order to consider other possible causes. These (N = 50), as listed in Table 1, included confirming whether the assessment is correct (7), guidance on additional information to obtain (9), and a request for other possible causes (broad differential) (34).

Elements Present in AI Responses

The AI responses were structurally consistent across cases. An explicit diagnostic list appeared in 41 of 50 responses (82%). Recommendations for further workup—additional history, focused examination, laboratory tests, or procedures—were present in all 50 responses (100%). Immediate treatment or initial management recommendations were also provided in all 50 responses (100%). Red flag features were explicitly identified in 26 cases (52%), and explicit management or follow-up concerns (such as return precautions, monitoring needs, or referral thresholds) were raised in 37 responses (74%) (Table 2).

Elements Present in Physician Responses

Physician responses showed a distinct pattern. An explicit diagnostic list or differential appeared in 12 of 50 responses (24%). Recommendations for next diagnostic steps were present in 42 responses (84%), representing the element most closely aligned with AI outputs. Immediate treatment recommendations appeared in 29 responses (58%). Explicit red flag features were identified in four cases (8%), and documented follow up or management concerns were found in nine responses (18%) (Table 3).

Discussion

In this case series from an African–U.S. telemedicine program, AI-generated responses from OpenEvidence and volunteer

Table 1. Reasons for telemedicine requests from Kenyan providers.

Reasons for telemedicine requests	Patients (*n*)	%
Confirm whether the assessment is correct	7	14
Guidance on what additional information to obtain	9	18
Request for other possible causes (broad differential)	34	68

Table 2. Elements present in AI responses.

Elements present	Patients (*n*)	%
List of possible diagnoses	4	82
Next steps suggested (history, labs, or procedures)	50	100
Immediate recommendations	50	100
Red flags identified	26	52
Concerns listed regarding management or follow-up	37	74

AI: artificial intelligence.

Table 3. Elements present in physician responses.

Physician responses	Number (*n*)	%
List of possible diagnoses	12	24
Next steps suggested (history, labs, or procedures)	42	84
Red flags identified	4	8
Concerns listed regarding management or follow-up	9	18
Immediate recommendations	29	58

physician responses exhibited markedly different structural patterns. AI outputs were systematically comprehensive, almost always including an explicit differential diagnosis, recommended next steps in workup, immediate management guidance, and frequent identification of red flag features and follow-up needs. Physician responses, by contrast, were more selective and focused, often emphasizing the most salient diagnostic or management issues without enumerating every potential diagnosis or safety concern.

These differences do not imply that AI is inherently superior to physicians or vice versa. Rather, they reflect distinct priorities. AI systems are designed to provide thorough, guideline-aligned, and internally consistent recommendations, which naturally lead to broad differentials and explicit safety checklists. Experienced clinicians, especially in resource-constrained settings, tend to prioritize what is feasible and immediately actionable, drawing on tacit knowledge about local capacity, patient context, and prior experience. As a result, safety concerns may be embedded implicitly in management plans rather than explicitly labeled as red flags, and differentials may be narrowed to the most likely or most consequential diagnoses.

In telehealth programs that support task shifting to clinical officers and other non-physician providers, both approaches have value. Structural completeness might help to ensure that important diagnostic possibilities and safety issues are not overlooked, particularly for less experienced clinicians. At the same time, advice that is too exhaustive or not calibrated to available resources can be difficult to implement, potentially overwhelming frontline teams. These findings therefore support a hybrid workflow in which AI serves as a baseline safety and completeness layer, while human consultants tailor and prioritize recommendations to the realities of Kenyan practice.

The results align with emerging literature on AI clinical decision support. This suggests that such tools can augment, rather than replace, clinician judgment by surfacing alternatives and providing explicit reasoning while clinicians retain responsibility for contextualization and final decisions. In a cross-border telemedicine context, using AI to pre-structure responses—highlighting red flags, listing differential diagnoses, and outlining generic workup options—could also reduce cognitive and documentation burden on volunteer physicians, allowing them to focus their limited time on local feasibility, trade-offs, and patient communication.

Limitations

The sample size was modest and drawn from a single telemedicine program, limiting generalizability. The analysis was descriptive and did not assess patient outcomes or provider-level effects such as confidence, satisfaction, or workflow impact. Physician responses were written for routine care and were not standardized for research, leading to heterogeneity in style and depth. OpenEvidence was the only AI platform evaluated; other tools or prompting strategies might yield different patterns of output. Finally, because all cases were de-identified and cross-sectional longitudinal effects of AI use on quality of care or safety incidents could not be examined.

Conclusions

Despite these limitations, this case series offers a concrete view of how AI-generated and physician-generated advice differ when applied to the same real-world telemedicine consultations in a low-resource setting.

The observed patterns suggest a natural division of labor: AI to ensure thoroughness and systematic safety checks, and physicians to supply contextual judgment, prioritization, and practical guidance grounded in local realities. For telehealth programs such as the Telemedicine Health and Empowerment Program, integrating AI into a supervised, human-in-the-loop workflow may enhance safety and consistency without displacing the relational and contextual expertise of human volunteers.

Funding

No external funding was received for this study.

Conflicts of Interest

The author declares no conflicts of interest.

Financial and Non-Financial Relationship and Activities

Not applicable.

Data Availability Statement (DAS), Data Sharing, Reproducibility, and Data Repositories

Due to the sensitive nature of clinical consultations, de-identified data is not publicly available but may be shared on reasonable request, subject to program approvals.

Application of AI-Generated Text or Related Technology

The following AI sites were used: Perplexity, Claude, and ChatGPT to help with formatting, gathering references, and checking for grammar, spelling, and word flow.

Acknowledgments

Not applicable.

References

1. Wootton R, Bonnardot L. Telemedicine in low-resource settings. Front Public Health. 2015;3:3. https://doi.org/10.3389/fpubh.2015.00003
2. Bonnardot L, Liu J, Wootton E, Amoros I, Olson D, Wong S, et al. The development of a multilingual tool for facilitating the primary–specialty care interface in low-resource settings: the MSF tele-expertise system. Front Public Health. 2014;2:126. https://doi.org/10.3389/fpubh.2014.00126
3. Kim EJ, Moretti ME, Kimathi AM, Chan SY, Wootton R. Use of provider-to-provider telemedicine in Kenya during the COVID-19 pandemic. Front Public Health. 2022;10:1028999. https://doi.org/10.3389/fpubh.2022.1028999
4. Totten AM, Womack DM, Griffin JC, McDonagh MS, Davis-O'Reilly C, Blazina I, et al. Telehealth-guided provider-to-provider communication to improve rural health: a systematic review. J Telemed Telecare. 2024;30(8):1209–29. https://doi.org/10.1177/1357633X221139892
5. Hurt RT, Stephenson CR, Gilman EA, Aakre CA, Croghan IT, Mundi MS, et al. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J Prim Care Community Health. 2025;16:21501319251332215. https://doi.org/10.1177/21501319251332215
6. Bagla P, Hanna J, Marthambadi B, Watkins S. Patterns of artificial intelligence use in clinical work by hospitalists: survey study. J Med Internet Res. 2026;28(1):e85973. https://doi.org/10.2196/85973

7. Ye J, He L, Beestrum M. Implications for implementation and adoption of telehealth in developing countries: a systematic review of China's practices and experiences. NPJ Digit Med. 2023;6(1):174. https://doi.org/10.1038/s41746-023-00908-6
8. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Network Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969
9. Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT, 3rd, Sheinberg M, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Network Open. 2025;8(2):e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637