Comparative Cost of Care Recommended by Large Language Models versus Treating Physicians

Authors: Dvir Aran, Ph.D. , Jeremy D. Goldhaber-Fiebert, Ph.D. , Ittai Many, M.D. , Anthony Nguyen, M.D. , Daniel Yang, M.D. , and Shahar Shelly, M.D. Author Info & Affiliations
Published October 15, 2026
NEJM AI 2026;3(11)
DOI: 10.1056/AIoa2600537
VOL. 3 NO. 11
Copyright © 2026

Abstract

BACKGROUND

Large language models (LLMs) are increasingly consulted for clinical decision support, yet the cost of services or treatment they recommend has not been systematically evaluated.

METHODS

We evaluated 24 AI systems on 200 deidentified outpatient and emergency department notes from MTSamples, a public collection of deidentified U.S.-format clinical notes contributed by practicing clinicians. Each system received the presentation recorded before the physician's assessment and wrote its own plan; the physician's documented plan was the comparator. Orders from both were extracted by three independent LLMs and priced at 2026 Medicare rates. The primary outcome was the per-visit total cost of recommended care: diagnostic tests plus specialist referrals. New medications were counted rather than priced. We also tested two prompts: one cost-aware prompt and one parsimonious-but-safe prompt.

RESULTS

Across the 20 general-purpose systems, AI-suggested care cost a mean of US$314 per visit, compared with US$97 for care recommended by treating physicians — a difference of US$216 per visit (95% confidence interval, US$177 to US$257; P<0.001) or 3.2 times the physicians' cost. AI system mean costs ranged from US$209 to US$599. The difference reflected US$105 in diagnostic tests and US$111 in specialist referrals; systems also recommended more new medications (1.14 vs. 0.59 per visit). The four specialized medical systems were no less costly (2.7× to 3.5×). Physicians

recorded no diagnostic testing in 66% of visits; on those visits, every AI system suggested testing, usually a stereotyped panel. Prompting the systems to be cost-aware (2.5× the physician) or parsimonious (1.4×) reduced but did not remove the excess.

CONCLUSIONS

The care these systems recommended cost substantially more than physicians' documented care, chiefly because they suggested testing at visits where the physician had ordered none. LLM outputs are suggestions, not plans of care; final decisions should remain with clinicians.