Analyzing the Diagnostic Capabilities of Claude 3.5 Sonnet on Complex Clinical Cases

Jack Litwin, MS-2[1], Daniel J. Parente, M.D., Ph.D.[1,2]
[1]The University of Kansas School of Medicine-Kansas City, Kansas City, Kansas
[2]The University of Kansas Medical Center, Kansas City, Kansas, Department of Family Medicine and Community Health

Received Aug. 28, 2025; Accepted for publication Sept. 10, 2025; Published online Sept. 11, 2025
https://doi.org/10.17161/kjm.vol18.24447

Introduction. Despite increased healthcare spending, the United States faces significant challenges in healthcare delivery and outcomes. Large language models (LLMs) have shown promise in medical applications, but their clinical diagnostic capabilities require further investigation. This study evaluates the diagnostic performance of Claude 3.5 Sonnet, a new large language model (LLM), in complex medical cases compared to traditional medical journal readers.

Methods. We analyzed 20 case challenges from the New England Journal of Medicine. Each case was presented to Claude 3.5 Sonnet with full text and corresponding images. Diagnostic accuracy was measured and compared between Claude 3.5 Sonnet and medical journal readers.

Results. Claude 3.5 Sonnet achieved an overall diagnostic accuracy of 49.5%, significantly higher than medical journal readers at 27.4% ($p = 0.042$). The AI model showed perfect accuracy (100%) in 9 cases and no accuracy (0%) in 10 cases, with one case at 90% accuracy. Reader accuracy ranged from 9% to 63% across cases.

Conclusions. Claude 3.5 Sonnet demonstrated significantly higher diagnostic accuracy compared to medical journal readers, though with notable variability in performance. These findings suggest potential utility for AI assistance in medical diagnosis, however further research comparing consistency and reliability of AI diagnostic capabilities to that of physicians is required.