Poster 15: How does ChatGPT perform on the Gynecologic Oncology and Critical Care PROLOG Exam?

Presenting Author: Eric Helm, MD, University of Colorado School of Medicine

Topic

Other: Artificial Intelligence in Medicine

Objectives

Artificial intelligence (AI) is rapidly integrating into daily technologies, and its use has been proposed as an adjunct in medical decision making. This study sought to assess the common AI tool ChatGPT's performance on board-level gynecologic oncology questions across different cognitive complexity levels and content domains.

Methods

We evaluated ChatGPT's performance on the 8th edition of PROLOG Gynecologic Oncology and Critical Care questions developed by the American College of Obstetricians and Gynecologists. Questions were categorized using Bloom's Taxonomy into three difficulty levels: Easy (Knowledge, Understand), Moderate (Apply), and Advanced (Analysis, Evaluate). Questions were further classified by content area: counseling, epidemiology and biostatistics, medical management, screening and diagnosis, and surgical management. Questions requiring image analysis (n=12) were excluded. Statistical analysis included chi-square tests for categorical comparisons and 95% confidence intervals for proportion estimates.

Results

136 questions were included for analysis. ChatGPT achieved an overall accuracy of 78.7% (107/136, 95% CI: 71.8%-85.6%). Performance declined significantly with increasing cognitive complexity: 88.0% on Easy questions (44/50, 95% CI: 79.0%-97.0%), 79.6% on Moderate questions (39/49, 95% CI: 68.3%-90.9%), and 64.9% on Advanced questions (24/37, 95% CI: 49.5%-80.3%) (p=0.033). Content area analysis revealed variable performance: 62.5% on counseling (5/8, 95% CI: 29.0%-96.1%), 100% on epidemiology (6/6, 95% CI: 100%), 82.5% on medical management (52/63, 95% CI: 73.2%-91.9%), 73.9% on screening/diagnosis (17/23, 95% CI: 56.0%-91.9%), and 75.0% on surgical management (27/36, 95% CI: 60.9%-89.2%) (p=0.046).

Conclusions

ChatGPT approached but did not reach the 80.0% passing threshold on the PROLOG exam. Performance was significantly influenced by cognitive complexity, with a 23.1% decline in accuracy from Easy to Advanced questions. Content analysis revealed worse performance in counseling scenarios, suggesting limitations when clinical judgment rather than factual knowledge is required. These findings highlight both the potential and limitations of AI in medical education and emphasize the need for establishing validation protocols before clinical implementation.