Evaluating the Quality of AI-Generated Health Information on Perinatal Psychopharmacology

Lydia J Kim[1], Kurt R Wharton[2]

[1]Oakland University William Beaumont School of Medicine, [2]Department of Obstetrics and Gynecology, Corewell Health William Beaumont University Hospital

Introduction

Recent surveys indicate that nearly half of consumers use generative AI for health-related inquiries,[1] driven largely by the convenience these platforms offer. As these tools gain widespread popularity, patients are increasingly turning to AI chatbots as convenient sources of medical information,[2] a trend with significant implications for patient safety. AI chatbots can produce incorrect information, warranting caution in medical practice.[2] Standardized frameworks for evaluating the quality of AI-generated health information remain underdeveloped. This study aims to evaluate and compare the quality of responses provided by ChatGPT, Gemini, and Claude to a clinically realistic patient query.

Aims and Objectives

1. To evaluate and compare the quality of health information generated by three publicly available AI platforms (ChatGPT, Gemini, and Claude) in response to a clinically realistic patient query about the safety of sertraline use during pregnancy and breastfeeding.
2. To identify specific domains of strength and weakness in AI-generated perinatal medication information across all three platforms.

Methods

Each of the three AI platforms - ChatGPT (OpenAI), Gemini (Google), and Claude (Anthropic) - received a single, identical prompt simulating a realistic patient query from a 30-year-old, 10-week pregnant first-time mother with a history of major depressive disorder asking about the safety of continuing sertraline throughout pregnancy and breastfeeding, including dose adjustment. Each platform's combined response was evaluated by one rater using the DISCERN instrument, producing item-level scores for Questions 1-16 and a total score ranging from 16 to 80.

Results

Claude received the highest total score (44), followed by Gemini (41) and ChatGPT (37). Using published DISCERN quality benchmarks, all three platforms fell within the "fair quality" range (39-50), with the exception of ChatGPT, which fell just below this threshold in the "poor quality" range (27-38).

Table 1. Total DISCERN Scores by Platform

Platform	Total DISCERN Score	Quality Classification
Claude	44	Fair
Gemini	41	Fair
ChatGPT	37	Poor

Conclusions

Universal deficiencies were identified across all platforms, particularly in acknowledging uncertainty in current evidence, presenting alternative treatment options, and directing patients to additional sources of support. These findings suggest that publicly available AI platforms should not be relied upon as standalone sources of perinatal medication information, and highlight the need for ongoing quality evaluation of AI-generated health content against validated consumer health information standards.

References

1. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence–generated health information in the United States: cross-sectional questionnaire study. *J Med Internet Res.* 2024;26:e55138. doi:10.2196/55138
2. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model [Preprint]. *Res Sq.* 2023. doi:10.21203/rs.3.rs-2566942/v1)

Acknowledgements

I would like to acknowledge the support provided by my mentor, Dr. Kurt Wharton. I am also grateful to Dr. Dwayne Baxa, my advisor, whose guidance was invaluable.