Artificial Intelligence Chatbot Knowledge on the Diagnosis and Treatment of Common Retinal Disorders

Arshia Eshtiaghi, MD, Andrew Mihalache, MD(C), Ryan Huang, MSc MD(C), Marko Popovic, MD MPH, Rajeev Muni, MD MSc FRCSC,

[1]Department of Ophthalmology and Vision Science, University of Toronto
[2]Temerty Faculty of Medicine, University of Toronto

Introduction: ChatGPT is a large language model that operates by predicting and generating text based on patterns learned from a diverse range of internet resources during its development. We aim to evaluate ChatGPT's responses to diagnostic and therapeutic questions on common retinal disorders.

Methods: In this cross-sectional study, we prompted ChatGPT with questions regarding diabetic retinopathy (DR), retinopathy of prematurity (ROP), age-related macular degeneration (AMD), epiretinal membrane (ERM), macular hole (MH), posterior vitreous detachment (PVD), rhegmatogenous retinal detachment (RRD), retinoschisis (RS), retinitis pigmentosa (RP), retinal artery occlusion (RAO), and retinal vein occlusion (RVO). Two retina specialists independently graded responses using a Likert scale ranging from 1 (unacceptable inaccuracies) to 5 (no inaccuracies). Our primary endpoint was the median grade given to ChatGPT-3.5 and ChatGPT-4's responses. Our secondary endpoints were differences between the two chatbot models in mean response time, length in characters and readability scores.

Results: ChatGPT-3.5 performed worst (median grade=3/5) on questions pertaining ROP, RS, or RVO and best (median grade=4/5) on questions pertaining to DR, AMD, ERM, PVD, RP, or RRD. ChatGPT-4 performed worst (median grade=4/5) on questions pertaining to ROP, MH, RS, RP, or RVO and best (median grade=4.5/5) on questions pertaining to DR, AMD, ERM, PVD, RRD, or RAO. ChatGPT-4 (81.8%) achieved a greater proportion of responses with a grade of at least 4/5 than ChatGPT-3.5 (54.5%; $p<0.01$). ChatGPT-4 took significantly longer to generate responses compared to ChatGPT-3.5 ($p<0.01$) and produced significantly longer responses ($p<0.01$). Evaluation using readability indices indicated that the responses of ChatGPT-4 tended to be more challenging to read than the responses of ChatGPT-3.5.

Conclusions: ChatGPT provides valuable responses on retinal disorders, although may sometimes lack nuance and have important omissions. Patients should appreciate the educational potential of chatbots in ophthalmology while approaching them with caution.