Abstract 129

COMPARING THE ABILITY OF LARGE LANGUAGE MODELS TO DIAGNOSE AND MANAGE COMMON MEDICAL RETINA AND VITREORETINAL PRESENTATIONS

Gupta A.*, Anastasi M., Mohite A., Nazarova D., Asaria R.

Royal Free London NHS Foundation Trust ~ London ~ United Kingdom

Advances in technology have resulted in integration of artificial intelligence (AI) in ophthalmology. As AI continues to develop, there is potential for use in diagnosis and management support. This study assesses the performance of the latest Chat-GPT models in answering questions related to retinal pathology images. The aim is to evaluate and compare the accuracy, relevance, and clinical usefulness of responses generated by GPT-4o, a personally trained ChatGPT model (the Retina Specialist), and GPT-o1 when answering retina-related clinical questions. We also assess the consistency of the models' responses across different retinal conditions considering their frequency of presentation, and to determine differences in grading responses between clinicians of different seniority.

This is an observational study to assess the ability of three Chat-GPT models, GPT-4o, a personally trained ChatGPT model (the Retina Specialist), and GPT-o1 to diagnose and manage 15 common retinal conditions based on imaging alone.
Cases and images were selected in line with the American Academy of Ophthalmology's Preferred Practice Pattern guidelines, and reviewed by a panel of 6 independent ophthalmologists. Each case was presented non-sequentially as new threads to prevent AI learning.
The 45 responses were assessed independently by 6 blinded UK-based consultant and fellow vitreoretinal and medical retina surgeons. Responses were presented in a random order using the Qualtrics survey system and graded using a standardised Likert scale score from 1 (significantly incorrect or irrelevant) to 5 (completely correct and directly relevant) based on accuracy and clinical applicability, clarity and conciseness, and humanity or best practice guidelines.

Initial analysis of 45 AI-generated responses shows variable performance between models. Across all conditions, GPT-4o and the Retina Specialist achieved higher mean Likert scores compared to GPT-o1; however, differences were modest. The mean scores were 4.1 (±0.5) for GPT-4o, 4.0 (±0.5) for the Retina Specialist, and 3.5 (±0.6) for GPT-o1. Although GPT-4o produced more accurate and clinically relevant answers overall, inconsistencies were noted.
Responses from GPT-o1 were more frequently graded as lacking sufficient detail or containing clinically ambiguous recommendations. Across all models, common retinal conditions such as diabetic macular oedema and neovascular age-related macular degeneration yielded higher scores compared to less common diagnoses.
Consultant graders consistently assigned lower scores than fellows, suggesting higher expectations regarding clarity, management appropriateness, and adherence to best practice guidelines.

These preliminary findings suggest that while GPT-based models demonstrate potential to assist in the interpretation of retinal imaging, significant limitations remain, particularly when addressing complex or less frequently encountered diseases. Further detailed analysis, including intergrader reliability and subgroup comparisons by disease category, is ongoing.