Interpretation of Clinical Retinal Images Using an Artificial Intelligence Chatbot

Andrew Mihalache[1], MD(C), Ryan S. Huang[1], MD(C), David Mikhail[1], MD(C),
Marko M. Popovic[2], MD MPH, Rajeev H. Muni[2] MD MSc FRCSC,

[1]Temerty of Medicine, University of Toronto
[2]Department of Ophthalmology and Visual Sciences, University of Toronto

Introduction: The subspecialty of retina is dependent on nuanced interpretations of multimodal imaging to ensure high diagnostic accuracy. This investigation aims to assess the performance of ChatGPT-4 in providing accurate diagnoses to multimodal retina cases from OCTCases, a medical education platform from the Department of Ophthalmology and Vision Sciences at the University of Toronto.

Methods: We prompted a custom chatbot with 69 retina teaching cases containing multimodal ophthalmic images, asking it to provide the most likely diagnosis. In a sensitivity analysis, we inputted increasing amounts of clinical information pertaining to each case until the chatbot achieved a correct diagnosis. Our primary outcome was the proportion of cases for which the chatbot was able to provide a correct diagnosis. Our secondary outcome was the chatbot's performance in relation to the amount of text-based information accompanying ophthalmic images. We performed multivariable logistic regressions on Stata v17.0 (StataCorp LLC, College Station, Texas) to investigate associations between the amount of text-based information inputted per prompt and the odds of the chatbot achieving a correct diagnosis, adjusting for the eye laterality of cases, number of ophthalmic images inputted, and imaging modalities.

Results: Across 69 retina cases collectively containing 139 ophthalmic images, the chatbot was able to provide a definitive, correct diagnosis for 35 (50.7%) cases. The chatbot needed variable amounts of clinical information to achieve a correct diagnosis, where the entire patient description as presented by OCTCases was required for a majority of correctly diagnosed cases (23/35 cases, 65.7%). Relative to when the chatbot was only prompted with a patient's age and sex, the chatbot achieved a higher odds of a correct diagnosis when prompted with an entire patient description (OR=10.1, 95%CI=[3.3, 30.3], $p<0.01$). Despite providing an incorrect diagnosis for 34 (49.3%) cases, the chatbot listed the correct diagnosis within its differential diagnosis for 7 (20.6%) of these incorrectly answered cases.

Conclusions: This custom chatbot was able to accurately diagnose approximately half of the retina cases requiring multimodal input, albeit relying heavily on text-based contextual information that accompanied ophthalmic images. The diagnostic ability of the chatbot in interpretation of multimodal imaging without text-based information is currently limited. The appropriate use of the chatbot in this setting is of utmost importance, given bioethical concerns.