Comparison of ChatGPT and Gemini in Surgical Planning for Orbitotomy

Jenny Ma[1], MD, Kenneth Chang[1], MD, Georges Nassrallah[1], MD, FRCSC,

[1]Department of Ophthalmology and Vision Sciences, University of Toronto

Introduction: There has been an increasing use of artificial intelligence to aid radiologic imaging interpretation and to complement clinical decision making. However, their role in ophthalmologic surgical decision making has not been assessed. This study aims to assess the ability of ChatGPT and Gemini to interpret diagnostic imaging reports and to recommend appropriate surgical approaches for patients undergoing orbitotomy.

Methods: We conducted a consecutive, retrospective case series of all adult orbitotomy cases from July 2021 to September 2023 of three oculoplastic surgeons at University of Toronto, Ontario, Canada. Thirty-four patients underwent an orbitotomy. For each patient, the computed tomography (CT) or magnetic resonance imaging (MRI) report was input into ChatGPT 3.5 and Gemini 1.0. A standardized script was used to ask four questions: 1) Top 3 differential diagnosis; 2) Single most likely diagnosis; 3) Most appropriate biopsy type (incisional vs. excisional); and 4) Recommended surgical approach to access the lesion. Our outcomes included the proportion of cases where the differential diagnosis included the final pathology diagnosis and where the recommended biopsy type and surgical approach matched the surgeon's operative choice.

Results: The analysis included 30 patients. Gemini only answered the script questions for a subset of 17 patients, and declined to give medical advice for the remainder. For this subset, ChatGPT and Gemini performed very similarly. The top 3 differential diagnoses based on CT or MRI report findings included the final pathology diagnosis in 71% of cases for both. The proposed most likely diagnosis matched the pathology diagnosis in 53% vs. 47% of cases (p=1.0), respectively. The suggested biopsy type (incisional vs. excisional) matched the surgeon's choice in 67% vs. 47% (p=0.25). When asked regarding the most appropriate surgical approach to access the lesion, their recommendation matched the surgeon's choice in 45% vs. 36% (p=1.0). Looking at only ChatGPT data for all 30 patients, its performance was: 1) 50%; 2) 38%; 3) 72%; and 4) 39%. For 5 patients, ChatGPT indicated neither type of biopsy as appropriate as the suspected diagnosis was inflammatory in nature. In two patients, it indicated neither biopsy was appropriate and suggested different biopsy techniques such as stereotactic biopsy or fine-needle aspiration biopsy.

Conclusions: ChatGPT and Gemini demonstrate potential in diagnostic imaging interpretation and in aiding preoperative surgical planning. However there remains limitations in their ability to accurately interpret radiologic imaging findings without clinical context and to select an appropriate surgical approach, illustrating the complexity and nuance of this decision.