B26

Ezanna Mesfin B.S. Advisor(s): Inci Dersu M.D.

Co-author(s):

Evaluating ChatGPT-4o's Reasoning in Ocular Injury Triage Using NEISS Data

This study evaluates ChatGPT-4o's reasoning process and effectiveness in triaging ocular injury cases by classifying them as emergent, urgent, or routine using data from the National Electronic Injury Surveillance System (NEISS). The goal is to analyze the reasoning types used by the model and their accuracies to enhance AI-assisted triage systems.We analyzed 2,145 ocular injury cases randomly sampled from the NEISS database, categorized into three triage levels: emergent, urgent, and routine. ChatGPT-4o was tasked to assign triage levels, recommend interventions, and provide reasoning for each decision. The model's reasoning was categorized into four types: 1) "Blunt trauma may cause delayed complications like retinal detachment," 2) "Chemical injuries can cause severe ocular damage and require immediate attention," 3) "Foreign body cases can cause vision-threatening complications if not treated urgently," and 4) "The injury appears non-urgent based on provided details." We evaluated the frequency and accuracy of each reasoning type and analyzed their association with triage categories. ChatGPT-4o predominantly used the reasoning "The injury appears non-urgent," in 1,395 cases with an accuracy of 51.1%. The "Foreign body" reasoning was used in 741 cases with a 76.7% accuracy rate, effectively identifying emergent cases but causing errors in urgent cases. The "Chemical injuries" reasoning achieved 100% accuracy but was underutilized, appearing in only three cases. The "Blunt trauma" reasoning appeared six times with a 50% accuracy rate. Overall, ChatGPT-4o demonstrated a 60% accuracy in triage classification, showing a bias toward routine classifications and underrepresentation of urgent cases. ChatGPT-4o shows potential for triaging ocular injuries, especially in identifying emergent cases. However, it struggles with nuanced reasoning for urgent classifications. Targeted retraining and data balancing could improve its reliability in clinical settings.