ChatGPT 4 Vision: Unveiling Its Educational Potential in Orthopaedic Trauma Cases

Arthur Pierre Drouaud, BS*; Carolina Stocchi, BS; Justin Evan Tang, BS; Jan Paul Szatkowski, MD; David Forsh, MD*

Purpose: We evaluated the performance of the novel ChatGPT 4 Vision (GPT-4V), with its capabilities of text and image interpretation, in orthopaedic trauma cases from OrthoBullets. This study aims to assess GPT-4V's image interpretation, diagnosis formulation, and patient management capabilities, shedding light on its potential as an educational tool for health-care professionals and patients.

Methods: 10 of the most popular orthopaedic trauma cases from Orthobullets were selected. Medical imaging and patient information were input into the GPT-4V chatbox, prompting the large language model to interpret the images, form a diagnosis, and guide responses to Orthobullet questions. Four board-certified orthopaedic trauma surgeons individually rated GPT-4V responses using a 5-point Likert scale (strongly disagree to strongly agree). Each of GPT-4V's answers were assessed for alignment with current medical knowledge and guidelines (accuracy), rationale and whether it is logical and understandable (rationale), relevancy to the specific case (relevance), and whether surgeons would trust the medical information provided (trustworthiness). Mean scores from surgeon ratings were calculated.

Results: In total, 10 clinical cases, comprising 97 questions, were analyzed (10 imaging, 35 management, 52 treatment). The surgeons assigned an average overall rating of 3.46/5.00 to GPT-4V's imaging response (with scores for accuracy at 3.28, rationale at 3.68, relevance at 3.75, and trustworthiness at 3.15). Management questions received an overall score of 3.76 (accuracy 3.61, rationale 3.84, relevance 4.01, trustworthiness 3.58), while treatment questions had an average overall score of 4.04 (accuracy 3.99, rationale 4.08, relevance 4.15, trustworthiness 3.93). Interrater agreement ranged from 0.11 to 0.0055 between surgeons due to Likert scale subjectivity, case complexity, and rater background training heterogeneity.

Conclusion: This is the first study evaluating GPT-4V's ability to interpret orthopaedic trauma imaging, develop personalized management, and offer treatment approaches. Surgeon ratings indicated moderate agreement. GPT- 4V performed less favorably in imaging compared to management and treatment suggesting that the utility of GPT-4V for interpreting radiographic images falls short compared to its performance in management and treatment approaches. Further advancements are warranted to optimize the educational utility in clinical scenarios.