Cracking the Ortho Code: Assessing ChatGPT's Proficiency on the Orthopaedic In-Training Examination with the Addition of Image Processing Capabilities

Kevin M. Posner[1],Cassandra Bakus[1], Grace Chester[1], Grace Basralian[1], Mallery Zeiman[1], Geoffrey R. O'Malley[1], Gregg R. Klein M.D.[2]

[1] Hackensack Meridian School of Medicine, Nutley, NJ, [2] Hackensack University Medical Center, Department of Orthopaedics, Hackensack, NJ

kevin.posner@hmhn.org

Disclosures: Kevin M. Posner (N), Cassandra Bakus (N), Grace Chester (N), Grace Basralian (N), Mallery Zeiman (N), Geoffrey R. O'Malley (N), Gregg R. Klein M.D. (3B; Zimmer Biomet. 7A; Zimmer Biomet.)

INTRODUCTION:The integration of artificial intelligence models like ChatGPT in healthcare is under scrutiny, with their efficacy yet to be conclusively determined in specific medical fields, including orthopaedic surgery. These systems are constantly upgrading and evolving allowing for the incorporation of new features, such as image analysis. The ability for image analysis allows for a more complete assessment of the possible implementation of AI into orthopaedic practice. This study assesses ChatGPT's performance in answering Orthopaedic In-Training Examination (OITE) questions, including those that include images, in comparison to the performance of orthopaedic surgery residents as well as the repeatability of the system's answer choices.

METHODS:Questions from the 2014, 2015, and 2021 AAOS OITE were screened for inclusion. Questions that required the use of a video were excluded, however, all questions that necessitated the use of images were included. Analysis of 746 questions yielding inclusion of 733 questions total. All questions without images were entered into ChatGPT 3.5 and 4.0 twice. Questions that necessitated the use of images were entered into ChatGPT 4.0, as this is the only version of the system that can analyze images. The responses were recorded and compared to AAOS's correct answers, evaluating the AI's accuracy and consistency.

RESULTS SECTION:Of the 733 questions included in final analysis, 358 questions did not necessitate image analysis by ChatGPT in order to provide an answer in comparison to 375 questions which did require image analysis. ChatGPT 4.0 performed significantly better on questions that did not require image analysis (70.39% vs 43.73%, $p<0.001$). When assessing the consistency of ChatGPT in regards to questions without images, there was no significant difference between the average number of correct responses on initial and second entry of questions for either the 3.5 or 4.0 systems ($p=0.342$; $p=0.12$). However, on average, ChatGPT 4.0 did perform significantly better on questions that did not require images when compared to ChatGPT 3.5 on the same set of questions (67.78% vs 52.20%, $p<0.001$).

DISCUSSION:While the use of artificial intelligence in medicine is an intriguing possibility, this evaluation demonstrates how a system such as ChatGPT still falls short. While ChatGPT clearly is able to provide answers to orthopedic based questions and the newest addition of image processing capabilities of ChatGPT 4.0 is promising, there is still a lack of accuracy that raises concern. ChatGPT 4.0 is an improvement on its predecessor, ChatGPT 3.5, with improved accuracy, however despite its newest feature, ChatGPT 4.0 is unable to answer questions accurately that necessitate the use of image/radiographic analysis. This further elucidates the idea that orthopedic surgery requires a high degree of clinical reasoning with the use of image analysis, a skill that ChatGPT still lacks. As AI technology evolves, ongoing research is vital to harness AI's potential effectively, ensuring it complements rather than attempts to replace the nuanced skills of healthcare professionals.

SIGNIFICANCE/CLINICAL RELEVANCE:This study is significant as it evaluates the current capabilities and boundaries of AI, particularly ChatGPT, in orthopedic surgery. The findings could be instrumental in guiding how quickly and how efficiently ChatGPT may be integrated into clinical practice and education as it aids in bridging knowledge gaps and creating more informed decision-making processes.