ChatGPT Performs at the Level of a Third-Year Orthopaedic Surgery Resident on the Orthopaedic In-training Examination

Diane Ghanem[1], Oscar Covarrubias[2], Micheal Raad[1], Dawn LaPorte[1], Babar Shafiq[1]
[1]The Johns Hopkins Hospital, Baltimore, MD [2]Johns Hopkins University, School of Medicine, Baltimore, MD
dghanem1@jh.edu

Disclosures: No relevant disclosures.

INTRODUCTION: Standardized exams have long been considered a cornerstone in measuring cognitive competency and academic achievement. Their fixed nature and predetermined scoring methods offer a consistent yardstick for gauging intellectual acumen across diverse demographics. Consequently, the performance of artificial intelligence (AI) in this context presents a rich, yet unexplored terrain for quantifying AI's understanding of complex cognitive tasks and simulating human-like problem-solving skills. Publicly available AI language models such as ChatGPT have demonstrated utility in text generation and even problem-solving when provided with clear instructions. Amidst this transformative shift, the aim of this study is to assess ChatGPT's performance on the orthopaedic surgery in-training examination (OITE).

METHODS: All 213 OITE 2021 web-based questions were retrieved from the AAOS-ResStudy website. Two independent reviewers copied and pasted the questions and response options into ChatGPT Plus (version 4.0) and recorded the generated answers. All media-containing questions were flagged and carefully examined. Twelve OITE media-containing questions that relied purely on images (clinical pictures, radiographs, MRIs, CT scans) and could not be rationalized from the clinical presentation were excluded. Cohen's Kappa coefficient was used to examine the agreement of ChatGPT-generated responses between reviewers. Descriptive statistics were used to summarize the performance (% correct) of ChatGPT Plus. The 2021 norm table was used to compare ChatGPT Plus' performance on the OITE to national orthopaedic surgery residents in that same year.

RESULTS: A total of 201 were evaluated by ChatGPT Plus. Excellent agreement was observed between raters for the 201 ChatGPT-generated responses, with a Cohen's Kappa coefficient of 0.947. 45.8% (92/201) were media-containing questions. ChatGPT had an average overall score of 61.2% (123/201). Its score was 64.2% (70/109) on non-media questions. When compared to the performance of all national orthopaedic surgery residents in 2021, ChatGPT Plus performed at the level of an average PGY3.

DISCUSSION: ChatGPT Plus is able to pass the OITE with a satisfactory overall score of 61.2%, ranking at the level of third-year orthopaedic surgery residents. More importantly, it provided logical reasoning and justifications that may help residents grasp evidence-based information and improve their understanding of OITE cases and general orthopaedic principles.

SIGNIFICANCE/CLINICAL RELEVANCE: With further improvements, AI language models, such as ChatGPT, may become valuable interactive learning tools in resident education, although further studies are still needed to examine their efficacy and impact on long-term learning and OITE/ABOS performance.