Is ChatGPT Ready for Prime Time? Assessing the Accuracy of Artificial Intelligence in Answering Common Arthroplasty Patient Questions

Jenna Alysse Bernstein, David C Landy, Kimberly Karin Tucker

INTRODUCTION: Artificial Intelligence (AI) has exploded in use in society over the past year, but it is still unclear how it will be incorporated into medicine and specifically into orthopaedic surgery. While much discussion has centered on using AI for surgical decision making, we infer that AI, in its current form, may be easily used in offsetting some of the patient questions being fielded by surgeons and their staff. In this study, we sought to investigate how accurately chatGPT answered questions commonly asked by arthroplasty patients and thus determine whether it would be a good resource for arthroplasty patients.

METHODS: Two fellowship trained arthroplasty surgeons assembled a list of commonly asked patient questions by using questions in patient handbooks and from doing an internet search engine query of "best questions to ask my surgeon before knee/hip replacement." ChatGPT was then queried two times, one time asking the questions as written, and the second time prompting the chatGPT to answer the questions "as an orthopaedic surgeon" answering patient questions. Each surgeon then evaluated the accuracy of each set of answers by rating them on a 1-4 scale; 1 inaccurate, 2 partially accurate/partially inaccurate, 3 accurate, but incomplete, 4 accurate and complete. Agreement was assessed between the two surgeons' evaluation of each set of chatGPT answers using Cohen's kappa. The association between question prompt and response accuracy was assessed using the Wilcoxon signed-rank test.

RESULTS:

Eighty questions were queried to chatGPT in each data set. In the set of questions without a prompt for chatGPT, there was substantial agreement between the surgeons' evaluations (kappa =0.68). In the questions with a prompt for chatGPT to answer "acting as an orthopaedic surgeon," there was fair agreement between the evaluations with kappa of 0.33. When assessing the quality of the chatGPT responses, 21 of 80 responses (26%) had an average grade of less than 3 when asked without a prompt, and 6 of 80 responses (8%) had an average grade of less than 3 when preceded by a prompt. Responses to questions with a prompt asking ChatGPT to "act as an orthopaedic surgeon" had higher ratings than those without this prompt (P=0.03).

DISCUSSION AND CONCLUSION: ChatGPT performed substantially better when appropriately prompted to answer questions "as an orthopaedic surgeon" answering patient questions. There was a 92% accuracy of answers to these questions. As patients become more adept at using AI, there may be a benefit in developing a chat bot that is able to answer basic patient questions around the time of surgery as many patients are using the internet and other resources to gather medical information. While not all the answers were accurate, importantly, there were no dangerous answers given as determined by the evaluating surgeons. There was some disagreement between evaluations by the surgeons, which may be related to some differences in opinions about answers to the questions. At this time, chatGPT may be used by some patients to answer basic perioperative questions, but it is likely not quite ready for primetime.

	Accuracy with Act like Orthopaedic Surgeon Prompt			
Accuracy with No Prompt	1	2	3	4
1	0	0	0	2
2	1	2	9	7
3	0	0	8	8
4	0	3	12	28

Wilcoxon Signed-Rank Test P Value = .03