The Educational Potential of ChatGPT: Assessing Large Language Model ChatGPT Responses to Common Patient Questions Regarding Total Knee Arthroplasty

Yasir AlShehri, Gerard Anthony Sheridan, Lisa Howard[1], Donald S Garbuz[2], Bassam A Masri[3], Michael Neufeld[3]
[1]VGH, [2]UNIVERSITY OF BRITISH COLUMBIA, [3]University of British Columbia

INTRODUCTION:

Artificial intelligence (AI) has gained significant popularity in the medical field and holds enormous potential to improve health outcomes. Machine learning protocols in Orthopaedics are becoming more widespread. ChatGPT, a large language model based on AI technology, has become mainstream in recent years. ChatGPT uses natural language processing and deep learning to process vast amounts of data and generate response with a natural conversational flow.

Effective patient education has been found to decrease patient anxiety, postoperative complications, and promote a faster recovery. Currently, the evidence on the use of Large Language Models (LLMs) such as ChatGPT for patient education is limited. The purpose of our study was to assess ChatGPT's ability to accurately and adequately answer frequently asked patient questions related to TKA, based on expert opinion from specialized surgeons in the field of TKA.

METHODS:

Ten questions were obtained from the frequently asked questions (FAQ) patient education portal of The American Association of Hip and Knee Surgeons (AAHKS) website (https://hipknee.aahks.org/total-knee-replacement/) . The questions were selected and approved by the authors by consensus. Most questions required minor modifications to clarify the context to TKA (e.g changing "How long will it last?" to "How long will my knee replacement last?"). Each question was inputted into ChatGPT 3.5 on January 29, 2024, without additional prompts or limitations on response length (Figure 1). No alterations were made to the responses, and a new session was initiated for each question to remove context from previous interactions. The responses were evaluated by four subspecialty arthroplasty surgeons.

Quality analysis was performed using a grading system based on ones used in previously published literature. The ChatGPT responses, upon review by the surgeons, were graded from A to D (Table 1). Following individual grading of all responses by each surgeon, a consensus was attempted to be reached. Response readability was assessed using the Flesch-Kincaid Reading Ease Score (FRES) and Grade Level (FKGL), two widespread tools for assessing readability. The FRES provides a score from 0 to 100, with higher scores indicating easier readability. For example, scores of 70-80 correlate with a 7th-grade reading level, while scores of 30-50 reflect college-level difficulty. The FKGL correlates with the educational level required to comprehend the text, ranging from elementary to college levels in the US educational system.

RESULTS:

The ten responses were graded as follows (Table 2); A (Correct and sufficient response): 20% (n=2); B (Correct but insufficient response): 40% (n=4); C (Response containing correct and incorrect information): 30% (n=3); D (Incorrect response): 0% (n=0). One response (10%) was not given a concensus grade due to lack of agreement between the grading surgeons. The mean FRES was 25.3 [SD± 8.0], which corresponds to a college graduate level, and ranged from 10.4 to 41.1. The mean FKGL was 15.4 [SD±1.5], ranging from 12.9 to 18.5, also corresponding to a college graduate reading level (Table 3). The ten questions, response grades , and analyses are presented below.

DISCUSSION AND CONCLUSION:

This study aimed to assess the adequacy and accuracy of ChatGPT's responses to common TKA-related questions. The results show that it can provide responses with overall satisfactory accuracy to these questions. However, inaccuracies were encountered, and there was variability in the surgeons' grading of the responses. A critical distinction in our study is that accuracy does not presume adequacy. While ChatGPT may provide reasonable accuracy, it is often inadequate. ChatGPT in its current state does not incorporate intelligence but rather works on pattern recognition and machine learning platforms. As such, it can masquerade as sufficient but can often provide incorrect information. Patients need to be educated on its limitations as it becomes a commonly utilized tool.

In conclusion, ChatGPT provides accurate responses to common patient questions about TKA but lacks the depth of in-person discussions with surgeons. As the interface improves over time, outcomes will likely improve. According to the results of this study, ChatGPT can be a useful tool for patients. Patients should be encouraged to use such tools to enhance their understanding and facilitate informed consent discussions with their surgeon.

You

When can I drive after a total knee replacement?

ChatGPT

The timing for driving after a total knee replacement surgery can vary depending on several factors, including the speed of your recovery, your surgeon's recommendations, and any regulations in your area. In general, most people can resume driving within 4 to 6 weeks after surgery, once they have regained sufficient strength, flexibility, and control in their knee, and are no longer taking strong pain medications that may impair their ability to drive safely. However, it's crucial to follow your surgeon's specific guidance and to wait until you feel comfortable and confident behind the wheel to ensure your safety and the safety of others on the road. Always consult with your healthcare provider before resuming driving after any surgical procedure.

Response grading system	
A	Correct and sufficient response
B	Correct but insufficient response
C	Response containing correct and incorrect information
D	Incorrect response

Table 1: Response grading system

Response Number	Response Quality Grade				
	Surgeon #1	Surgeon #2	Surgeon #3	Surgeon #4	**Consensus**
1	A	B	A	A	**A**
2	B	C	C	A	NC
3	B	A	B	B	**B**
4	B	B	C	C	**C**
5	B	B	B	A	**B**
6	C	C	C	C	**C**
7	C	C	C	B	**C**
8	A	A	A	A	**A**
9	A	B	B	B	**B**
10	B	B	B	A	**B**

Table 2: Individual surgeon response grades and consensus grades (if reached). NC: No consensus

Response Number	Flesch-Kincaid Reading Ease Score	Flesch-Kincaid Grade Level
	100.00–91.00: 5th grade **90.0–81.0: 6th grade** **80.0–71.0: 7th grade** **70.0–65.0: 8th & 9th grade** **60.0–51.0: 10th to 12th grade** **50.0–31.0: College** **30.0–0: College graduate.**	**0-3: Kindergarten** **3.1-6: Elementary** **6.1-9: Middle school** **9.1-12: High school** **12.1-15: College** **15.1-18: College graduate**
1	24.5	15.1
2	35.1	13.9
3	35.5	14.7
4	16.9	16.0
5	41.1	12.9
6	24.2	15.3
7	10.4	18.5
8	24.3	17.3
9	30.8	14.8
10	22.5	15.9

Table 3: Flesch-Kincaid Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL) for the responses.