Assessing ChatGPT's Orthopedic In-Service Training Exam Performance and Applicability in the Field

Neil Jain[1], Caleb Gottlich[1], John Fisher[1], Dominic Campano[1], Travis Winston[1]
[1]Texas Tech University Health Sciences Center, Lubbock, TX
Njain5466@gmail.com

Disclosures: Neil Jain (None), Caleb Gottlich (None), John Fisher (None), Dominic Campano (None), Travis Winston (None)

INTRODUCTION:
ChatGPT has gained widespread attention for its ability to understand and provide human like responses to inputs. However, few works have focused on its use in Orthopedics. This study assessed ChatGPT's performance on the Orthopedic In-Service Training Exam (OITE) and evaluated its decision-making process to determine whether adoption as a resource in the field is practical.

METHODS:
ChatGPT's performance on three OITE exams was evaluated through inputting multiple choice questions. Questions were classified by their orthopedic subject area. Yearly OITE Technical Reports were used to gauge scores against resident physicians. ChatGPT's rationales were compared with testmaker explanations using six different groups denoting answer accuracy and logic consistency. Variables were analyzed using contingency table construction and chi-squared analyses.

RESULTS SECTION:
Of 635 questions, 360 were useable as inputs (56.7%). ChatGPT scored 55.8%, 47.7%, and 54% for the years 2020, 2021, and 2022, respectively. Of 190 correct outputs, 179 provided a consistent logic (94.2%). Of 170 incorrect outputs, 133 provided an inconsistent logic (78.2%). Significant associations were found between test topic and correct answer (p = 0.011), and type of logic used and tested topic (p = <0.001). Basic Science and Sports had adjusted residuals greater than 1.96. Basic Science and correct, no logic; basic science and incorrect, inconsistent logic; sports and correct, no logic; and sports and incorrect, inconsistent logic; had adjusted residuals greater than 1.96.

DISCUSSION:
ChatGPT performed around the level of a PGY-1 resident physician. When answering correctly, it displayed congruent reasoning with testmakers. When answering incorrectly, it exhibited some understanding of the correct answer. It outperformed in Basic Science and Sports, likely due to its ability to output rote facts. These findings and ChatGPT's inability to interpret radiographic imaging suggest that it lacks the fundamental capabilities to be a comprehensive tool in Orthopedic Surgery at this time.

SIGNIFICANCE/CLINICAL RELEVANCE:
ChatGPT represents a pinnacle of human achievement in the field of artificial intelligence with the potential to serve many educational functions due to its ability to self-improve using reinforcement learning techniques. This study demonstrates that it lacks the technical abilities to be a comprehensive tool in Orthopedic Surgery, but it may continue to have utility as a perioperative educational resource due to the high-quality sources in its trained data sets.

IMAGES AND TABLES:

Question Type * Logic (CC = correct consistent, CIC = correct inconsistent, IC = incorrect consistent, IIC = incorrect, inconsistent, CN = correct, no reasoning, IN = incorrect, no reasoning) Crosstabulation

Count

		Logic (CC = correct consistent, CIC = correct inconsistent, IC = incorrect consistent, IIC = incorrect, inconsistent, CN = correct, no reasoning, IN = incorrect, no reasoning)						
		CC	CIC	CN	IC	IIC	IN	Total
Question Type	Basic Science	14	0	5	1	2	0	22
	Foot and Ankle	14	0	0	3	15	1	33
	Hand and Wrist	9	1	0	2	14	1	27
	Hip and Knee	23	0	0	3	19	1	46
	Oncology	12	1	0	2	4	0	19
	Pediatrics	16	1	0	3	12	0	32
	Practice Management	18	0	0	1	12	1	32
	Shoulder and Elbow	17	0	0	4	17	2	40
	Spine	18	1	1	5	14	1	40
	Sports	18	1	0	3	4	0	26
	Trauma	20	0	0	2	20	1	43
Total		179	5	6	29	133	8	360

Table 1: Question Type and Logic Group Distribution

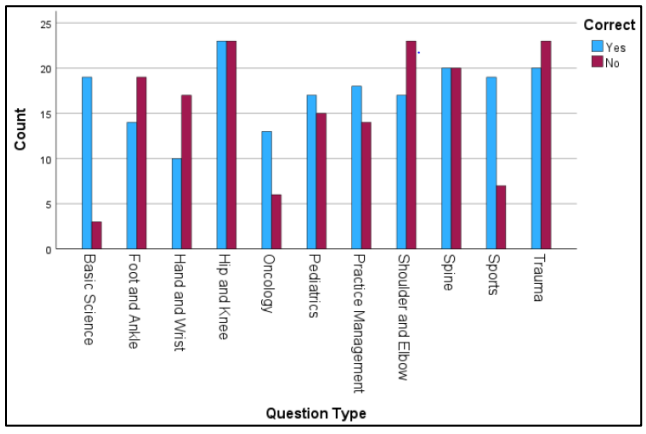

Figure 1: Correct or incorrect answers categorized by subject type

		Year			
		2020	2021	2022	Total
Correct	No	57	56	57	170
	Yes	72	51	67	190
Total		129	107	124	360

Table 2: Total questions answered correct on incorrect categorized by year