A Comparison Between the Efficacy of ChatGPT, Grok, and Claude Sonnet in Analyzing Common Sports-Related Radiologic Imaging

Michael Lane Moore, Ahmad R Alhankawi, Collin Braithwaite, Alex Miguel Holle, Anikar Chhabra

INTRODUCTION: The use of artificial intelligence (AI) has been a subject of major interest in the field of medicine. Previous studies have shown AI's powerful capabilities in the accuracy of medical inquiries, but a lack of data is published using radiologic images. The purpose of this paper is two-fold: (1) to assess the general ability for AI to diagnose common sport-related pathologies using radiologic imaging, and (2) to compare ChatGPT with two of its competitors, Grok and Claude Sonnet.

METHODS: ChatGPT 4.0, Grok 2, and Claude 3.5 Sonnet were utilized. Five common orthopedic sports pathologies were chosen: anterior cruciate ligament (ACL) tears, posterior cruciate ligament (PCL) tears, meniscal tears, chondral pathologies, and rotator cuff tears. Fifty images representing each pathology were randomly collected from a radiologic imaging database, when possible, which included radiographic images, computed tomography (CT), and magnetic resonance imaging (MRI). Normal images were collected that corresponded to each diagnostic category. A new query was used for each image to assess a likely diagnosis. Receiver operator characteristic curves and area under the curve values were calculated to assess the accuracy of each AI platform.

RESULTS: Regarding diseased images, ChatGPT, Grok and Claude Sonnet accurately identified the pathology in 23%, 16%, and 17% of images, respectively. ChatGPT and Grok were most accurate at identifying meniscus pathologies (ChatGPT: 48%, Grok: 42%), while Claude Sonnet was most accurate at identifying ACL pathologies (31%). The AUC for ChatGPT, Grok, and Claude Sonnet was 0.21, 0.16, and 0.16, respectively. There were no differences in performance between the three platforms overall or within any of the diagnostic categories. ChatGPT and Claude Sonnet had a statistically higher rate of correctly diagnosing diseased MRI images (ChatGPT $p<0.001$, Claude 3.5 Sonnet $p=0.003$) while Grok showed no differences in accuracy among imaging modalities ($p=0.088$)

DISCUSSION AND CONCLUSION: Open-Sourced AI platforms ChatGPT, Grok and Claude Sonnet had a diagnostic accuracy of less than 25% for common orthopedic sports pathologies, and had AUC values well under 0.5, indicating very low accuracy. Although ChatGPT generally operated better than Grok and Claude Sonnet, the overall results were unremarkable. We do not recommend its current use in image-based diagnoses.

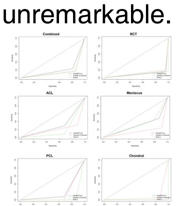

Figure 1: ROC curves generated for the overall combined number of images tested by ChatGPT 4.0, Grok 2, and Claude 3.5 Sonnet as well as the images ran by diagnostic category.

AI Platform	Total	ACL	PCL	Meniscus	Chondral	Rotator Cuff
ChatGPT 4.0						
# Correct	50	14	4	24	0	8
% Correct	23.3	28.6	8	48	0	16
Grok 2						
# Correct	34	4	0	21	1	8
% Correct	15.8	8.2	0	42	6.2	16
Claude 3.5 Sonnet						
# Correct	37	15	0	12	0	10
% Correct	17.2	30.6	0	24	0	20

Table 1: Number and percentage of diseased images correctly diagnosed by ChatGPT 4.0, Grok 2, and Claude 3.5 Sonnet.

Marker	AUC	Sensitivity	Specificity	PPV	NPV
ChatGPT					
Overall	0.21	0.23	0.19	0.38	0.1
ACL	0.31	0.28	0.32	0.45	0.18
PCL	0.2	0.08	0.32	0.21	0.13
Meniscus	0.31	0.48	0.13	0.55	0.1
Chondral	0	0	0	0	0
RCT	0.1	0.16	0.043	0.27	0.023
Grok 2					
Overall	0.16	0.16	0.16	0.29	0.081
ACL	0.2	0.08	0.32	0.19	0.15
PCL	0.068	0	0.14	0	0.057
Meniscus	0.32	0.42	0.22	0.54	0.15
Chondral	0.031	0.063	0	0.13	0
RCT	0.08	0.16	0	0.26	0
Claude 3.5 Sonnet					
Overall	0.16	0.17	0.14	0.3	0.073
ACL	0.22	0.3	0.12	0.41	0.08
PCL	0.16	0	0.32	0	0.12
Meniscus	0.16	0.24	0.087	0.36	0.05
Chondral	0.071	0	0.14	0	0.059
RCT	0.12	0.2	0.043	0.31	0.024

Table 2: AUC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) data for ChatGPT 4.0, Grok 2, and Claude 3.5 Sonnet.

Table 1. Demographics and Comorbidities Compared Between Cohorts

Demographics	Control (n=7032)	PTSD (n = 754)	P-value
Age (SD)	62.1 (9.0)	62.7 (8.7)	0.067
Male	3423 (48.7)	369 (48.9)	0.92
Female	3609 (51.3)	385 (51.1)	
CCI (SD)	2.8 (2.7)	2.9 (2.2)	0.22
Tobacco Use	4556 (64.8)	498 (66.0)	0.52
Obesity	4189 (59.6)	446 (59.2)	0.85
Alcohol Use	1446 (20.6)	181 (24.0)	**0.031**
Diabetes	3296 (46.9)	329 (43.6)	0.32
Depression	6408 (91.1)	688 (91.2)	0.97
Hypertension	6207 (88.3)	665 (88.2)	0.99
CKD	1661 (23.6)	186 (24.7)	0.55
Liver Disease	1935 (27.5)	226 (30.0)	0.16
SSRI Prescription	4485 (63.8)	474 (62.9)	0.65

*Reported as n (%) unless otherwise specified. PTSD, post-traumatic stress disorder; SD, standard deviation; CCI, Charlson Comorbidity Index; CKD, chronic kidney disease; SSRI, selective serotonin reuptake inhibitor. Bold indicates statistical significance ($p<0.05$).

Table 2. 90-Day Complications Compared Between Cohorts

Complication	Control (n=7032)	PTSD (n=754)	Odds Ratio	95% Confidence Interval	P-value
Myocardial Infarction	22 (0.3%)	5 (0.7%)	2.12	0.80-5.63	0.13
Pneumonia	110 (1.6%)	10 (1.3%)	0.85	0.44-1.62	0.62
Pulmonary Embolism	19 (0.3%)	1 (0.1%)	0.49	0.02-2.37	0.49
Deep Vein Thrombosis	46 (0.7%)	5 (0.7%)	1.01	0.35-2.32	0.98
Urinary Tract Infection	239 (3.4%)	32 (4.2%)	1.26	0.85-1.81	0.23
Blood Transfusion	49 (0.7%)	9 (1.2%)	1.53	0.68-3.06	0.27
Wound Dehiscence	22 (0.3%)	2 (0.3%)	0.85	0.14-2.88	0.82
Acute Kidney Injury	121 (1.7%)	22 (2.9%)	**1.72**	**1.06-2.66**	**0.022**
Sepsis	32 (0.5%)	2 (0.3%)	0.58	0.09-1.92	0.46
ED Visits	956 (13.6%)	152 (20.2%)	**1.60**	**1.32-1.94**	**<0.001**
Readmission	255 (3.6%)	37 (4.9%)	1.37	0.95-1.93	0.080

*Reported as n (%) unless otherwise specified. PTSD, post-traumatic stress disorder. Bold indicates statistical significance ($p<0.05$).

Table 3. 2-Year Complications Compared Between Cohorts

Complication	Control (n=7032)	PTSD (n=754)	Odds Ratio	95% Confidence Interval	P-value
Revision	171 (2.4%)	13 (1.7%)	0.70	0.40-1.24	0.23
Periprosthetic Joint Infection	128 (1.8%)	14 (1.9%)	1.02	0.56-1.72	0.94
Dislocation	117 (1.7%)	11 (1.5%)	0.88	0.44-1.56	0.67
Aseptic Loosening	214 (3.0%)	15 (2.0%)	0.65	0.37-1.06	0.11
Periprosthetic Fracture	35 (0.5%)	4 (0.5%)	1.07	0.32-2.68	0.90
Postoperative Stiffness	696 (9.9%)	95 (12.6%)	**1.31**	**1.04-1.64**	**0.020**

*Reported as n (%) unless otherwise specified. PTSD, post-traumatic stress disorder Bold indicates statistical significance ($p<0.05$).