Comparing Appropriate Use Criteria with ChatGPT-4 Recommendations for Treatment of Distal Radius Fractures

Kareem Mohamed, Akiro H Duey[1], Alexander Yu, Christoph Alexander Schroen[1], Jamie Kator[2], Michael R Hausman[3]
[1]Icahn School of Medicine At Mount Sinai, [2]Mount Sinai Orthopedics, [3]Mt Sinai Med Ctr

INTRODUCTION: The American Academy of Orthopaedic Surgeons (AAOS) appropriate use criteria (AUC) for managing distal radius fractures was created to guide treatment decisions based on expert panel vote. The goal of this study is to evaluate the accuracy of Chat Generative Pre-trained Transformer-4 (ChatGPT-4.0) by comparing its appropriateness scores for distal radius fracture treatment with that of the AUC.

METHODS: The AUC's patient scenarios were based on indication categories including Association of Osteosynthesis/Orthopaedic Trauma Association (AO/OTA) fracture type (type A, B or C), mechanism of injury (high or low-energy fracture), pre-injury activity level, patient health (ASA 1-2-3 or 4), and associated injuries (median neuropathy, Gustilo Anderson type I or II open fracture, Gustilo Anderson type III open fracture, other multi-trauma injury, or no associated injuries). Treatment options included percutaneous pinning, spanning external fixation, volar locking plate, dorsal plate fragment specific fixation, dorsal spanning bridge, intramedullary nail, immobilization without reduction, and reduction and immobilization. A panel of orthopedic surgeons from the AAOS voted for each treatment option given a patient scenario with its indications. A score from 7-9 indicates "Appropriate", 4-6 indicates "May Be Appropriate", and 1-3 indicates "Rarely Appropriate." For each set of indications evaluated by the AAOS panel, ChatGPT-4 was prompted to assign a rating for each treatment option. The AAOS ratings were subtracted from the ChatGPT-4 ratings to calculate the error. Mean error, mean absolute error, and mean squared error were calculated. Spearman correlation was used to determine statistical significance (α <.05). Each response was then categorized into the three appropriateness categories to find the percentage of overlap between AAOS and ChatGPT-4.

RESULTS: A total of 240 patient scenarios were evaluated for each of the 9 treatment options, providing a total of 2160 paired scores. Comparing ChatGPT-4 with AUC scores, The mean squared error was 3.9 ± 4.4 for percutaneous pinning, 6.7 ± 10.4 for spanning external fixation, 2.3 ± 5.1 for volar locking plate, 7.4 ± 7.3 for dorsal plate, 3.8 ± 7.5 for fragment-specific fixation, 4.5 ± 6.8 for dorsal spanning bridge, 3.6 ± 4.3 for intramedullary nail, 2.9 ± 10.1 for immobilization without reduction, and 14.4 ± 13 for reduction and immobilization (Table 1). Spearman correlation testing found that there was a significant positive correlation for volar locking plate (.14, P=.033), and dorsal spanning bridge (.14, P=.027) (Table 2). When AAOS and ChatGPT-4's ratings were grouped into "Appropriate," "May be Appropriate," or "Not Appropriate," the percentage overlap ranged widely, with 87.50% overlap in ratings for volar locking plate, 93.75% for immobilization without reduction, and 20.00% for reduction and immobilization (Table 3, Figure 1).

DISCUSSION AND CONCLUSION: Based on these findings, ChatGPT-4 is not able to reliably predict appropriate clinical management of distal radius fractures when compared to the AUC. Though there was relative concordance on volar plating and immobilization alone, ChatGPT-4 was more likely to trend towards conservative treatment modalities. As AI driven tools become more prominent and accessible, patients are able to seek medical counseling in ChatGPT-4. However, our study concludes that ChatGPT may inappropriately recommend nonoperative management when compared to validated AUC.

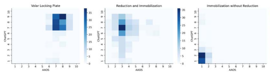

Figure 1. Heatmaps comparing ChatGPT and AAOS ratings of the appropriateness of volar locking plate,, reduction and immobilization and immobilization without reduction for 240 different patient scenarios.

Treatment	Mean Error	Mean Absolute Error	Mean Squared Error
Percutaneous Pinning	-1.3 ± 1.5	1.6 ± 1.6	3.9 ± 4.4
Spanning External Fixation	-1.5 ± 2.1	1.8 ± 1.9	6.7 ± 10.4
Volar Locking Plate	0.2 ± 1.5	1.1 ± 1.1	2.3 ± 5.1
Dorsal Plate	-2.2 ± 1.6	2.3 ± 1.4	7.4 ± 7.3
Fragment Specific Fixation	-0.7 ± 1.8	1.4 ± 1.3	3.8 ± 7.5
Dorsal Spanning Bridge	-1.4 ± 1.5	1.6 ± 1.4	4.5 ± 6.8
Intramedullary Nail	-0.5 ± 1.8	1.5 ± 1.1	3.6 ± 4.3
Immobilization without Reduction	0.8 ± 1.5	0.9 ± 1.4	2.9 ± 10.1
Reduction and Immobilization	3.1 ± 2.2	3.3 ± 1.9	14.4 ± 13

Table 1. Mean error, mean absolute error, and mean squared error between AAOS and ChatGPT scores for distal radius fracture treatment options.

Treatment	Spearman's rho	P-value
Percutaneous Pinning	0.09	0.156
Spanning External Fixation	-0.04	0.489
Volar Locking Plate	0.14	0.033
Dorsal Plate	-0.1	0.132
Fragment Specific Fixation	-0.03	0.657
Dorsal Spanning Bridge	0.14	0.027
Intramedullary Nail	0	0.942
Immobilization without Reduction	0.09	0.172
Reduction and Immobilization	-0.04	0.587

Table 2. Spearman correlation coefficients and p-values comparing AAOS and ChatGPT scores for distal radius fracture treatment options.

Treatment	Overlap Count	Total Cases	Percentage Overlap
Percutaneous Pinning	131	240	54.58%
Spanning External Fixation	112	240	46.67%
Volar Locking Plate	210	240	87.50%
Dorsal Plate	65	240	27.08%
Fragment Specific Fixation	117	240	48.75%
Dorsal Spanning Bridge/Wrist Plate	126	240	52.50%
Intramedullary Nail	115	240	47.92%
Immobilization without Reduction	225	240	93.75%
Reduction and Immobilization	48	240	20.00%

Table 3. Overlap of AAOS and ChatGPT ratings when grouped into "Appropriate," "May be Appropriate," or "Not Appropriate."