UC Davis
Orthopaedic Surgery

Title
Can AI Match Clinicians?Assessing Inter-rater Reliability in Fracture Identification

Permalink
https://escholarship.org/uc/item/42g182m2

Authors
Walker, Aliyah
Smith, J.B.
Simister, Samuel K
et al.

Publication Date
2025-04-01

Data Availability
The data associated with this publication are not available for this reason: NA

eScholarship.org Powered by the California Digital Library
University of California

Can AI Match Clinicians? Assessing Inter-rater Reliability in Fracture Identification

Aliyah N. Walker[2], J.B. Smith[1], Samuel K. Simister[1], Om Patel[2], Soham Choudhary[2], Michael Seidu[2], David Dallas-Orr[1], Shannon Tse[1], Hania Shahzad[1], Augustine Saiz[1], Zachary C. Lum[1]

[1]Department of Orthopedic Surgery, University of California Davis Health, Sacramento, CA, USA [2]University of California, Davis School of Medicine, Sacramento, CA, USA

Introduction

- In the US, approximately 275 million radiographs are performed each year
- This study compares the inter-rater reliability of a **conversational artificial intelligence model** to that of orthopaedic surgery **attendings and residents** in identifying **fractures on upper extremity (UE) and lower extremity (LE) radiographs**

Methods

- **84 radiographs of various fracture patterns** were presented to ChatGPT (Figure 1)
- Two orthopaedic surgery residents and two attending orthopaedic surgeons evaluated the same radiographs
- Radiograph reports between groups were compared according to **view, location, fracture type, and AO classification**
- **Inter-rater reliability (IRR) was evaluated** for the following groups using Fleiss' Kappa values:
  - All Raters Combined (ARR)
  - AI vs. Residents (AIR)
  - AI vs. Attendings (AIA)
  - Attendings vs. Residents (AR)

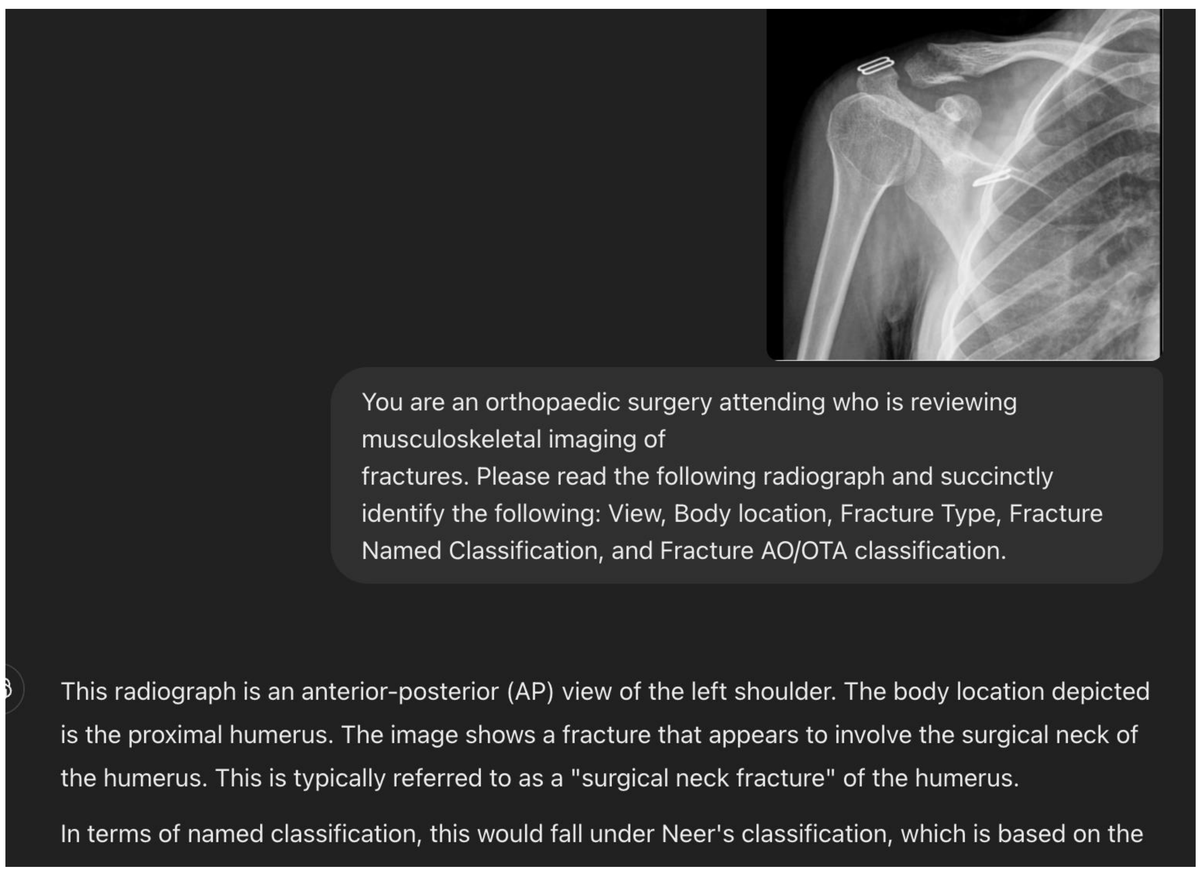

Figure 1: Screenshot of the ChatGPT prompt and output

Results / Discussion

Category	Comparison	Fleiss Kappa Value	P-Value	Strength of Agreement
View	All Raters Combined	0.488	$p < 0.001$	Moderate
	AI vs. Residents	0.404	$p < 0.001$	Fair
	AI vs. Attendings	0.37	$p < 0.001$	Fair
	Attendings vs. Residents	0.624	$p < 0.001$	Substantial
Location	All Raters Combined	0.767	$p < 0.001$	Substantial
	AI vs. Residents	0.708	$p < 0.001$	Substantial
	AI vs. Attendings	0.655	$p < 0.001$	Substantial
	Attendings vs. Residents	0.896	$p < 0.001$	**Almost Perfect**
Fracture Type	All Raters Combined	0.711	$p < 0.001$	Substantial
	AI vs. Residents	0.546	$p < 0.001$	Moderate
	AI vs. Attendings	0.563	$p < 0.001$	Moderate
	Attendings vs. Residents	0.948	$p < 0.001$	**Almost Perfect**
AO Classification	All Raters Combined	0.169	$p < 0.001$	Slight
	AI vs. Residents	-0.062	$p < 0.001$	Poor
	AI vs. Attendings	0.159	$p < 0.001$	Slight
	Attendings vs. Residents	0.257	$p < 0.001$	Fair

Table 1: IRR for **Upper Extremity Fracture** Identification

Category	Comparison	Fleiss Kappa Value	P-Value	Strength of Agreement
View	All Raters Combined	0.458	$p < 0.001$	Moderate
	AI vs. Residents	0.39	$p < 0.001$	Fair
	AI vs. Attendings	0.309	$p < 0.001$	Fair
	Attendings vs. Residents	0.654	$p < 0.001$	Substantial
Location	All Raters Combined	0.884	$p < 0.001$	**Almost Perfect**
	AI vs. Residents	0.909	$p < 0.001$	**Almost Perfect**
	AI vs. Attendings	0.834	$p < 0.001$	**Almost Perfect**
	Attendings vs. Residents	0.912	$p < 0.001$	**Almost Perfect**
Fracture Type	All Raters Combined	0.724	$p < 0.001$	Substantial
	AI vs. Residents	0.697	$p < 0.001$	Substantial
	AI vs. Attendings	0.58	$p < 0.001$	Moderate
	Attendings vs. Residents	0.859	$p < 0.001$	**Almost Perfect**
AO Classification	All Raters Combined	0.467	$p < 0.001$	Moderate
	AI vs. Residents	0.455	$P = 0.051$	Moderate
	AI vs. Attendings	0.418	$p < 0.001$	Moderate
	Attendings vs. Residents	0.517	$p < 0.001$	Moderate

Table 2: IRR for **Lower Extremity Fracture** Identification

Conclusions

- Although showing promise with identifying basic fracture features, **AI is not yet at a level comparable to experts**, especially with more nuanced interpretations
- Their use is more effective as diagnostic adjuncts rather than replacement for the judgment of trained clinicians