ORIGINAL PAPER/PŮVODNÍ PRÁCE

Diagnostic Performance of Artificial Intelligence (AI) Chatbot Compared to Orthopedic Trauma Surgeons in Evaluating Indication for Surgery for Isolated Lateral Malleolar Fractures: a Retrospective Study

Diagnostický výkon chatbota s umělou inteligencí (AI) ve srovnání s ortopedickými traumatology při hodnocení indikace k operaci izolovaných zlomenin laterálního malleolu: retrospektivní studie

MARIA OULIANSKI, RAMI MOSHEIFF, DANA AVRAHAM, OMER BEN YEHUDA, YORAM WEIL, MAHMOUD JAMMAL

Hadassah Medical Center of the Hebrew University, Jerusalem, Israel

Dedicated to the 70th anniversary of the birth of Prof. Martin Krbec, MD, CSc.

Corresponding author:

Jammal Mahmoud, MSc, MD
Kiryat Hadassah
POB 12000
Jerusalem, 91120, Israel

Jammalm@hadassah.org.il

Oulianski M, Mosheiff R, Avraham D, Ben-Yehuda O, Weil Y, Jammal M. Diagnostic Performance of Artificial Intelligence (AI) Chatbot Compared to Orthopedic Trauma Surgeons in Evaluating Indication for Surgery for Isolated Lateal Malleolar Fractures: a Retrospective Study. Acta Chir Orthop Traumatol Cech. 2026;93:133–139.

ABSTRACT

Purpose of the study

The indication for surgery for isolated lateral malleolar fracture (AO/OTA 44B1) is debatable and in many cases, relies upon radiographic assessment of fracture stability. Artificial intelligence chatbots with visual analysis capabilities offer a potential attribute for radiographic assessment. This study compared the diagnostic performance of a commercially available AI chatbot with that of three fellowship-trained orthopedic trauma surgeons in evaluating equivocal isolated lateral malleolar fractures.

Material and methods

A retrospective study. 50 patients with isolated lateral malleolar injury at the level of the syndesmosis (AO/OTA 44B1) were evaluated by three blinded fellowship-trained orthopedic trauma surgeons. Each rater measured standardized radiographic ankle parameters, medial clear space, tibiofibular clear space, and tibiofibular overlap, on anteroposterior and mortise views and determined a surgical versus nonoperative treatment recommendation. Subsequently, the same sets of radiographs were independently evaluated by an AI chatbot (Claude, Anthropic). The observers and AI decisions were compared to the actual outcome of the patients (operative vs. nonoperatives).

Results

All raters recommended surgery at lower rates (34.0–46.0%) than the actual operative rate (56.0%). The difference in outcomes between the actual treatment and the observers varied and ranged between 67.3–86% with the AI within the same ranges. The AI's radiographic measurements differed systematically from all surgeons across five of six parameters. Inter-rater agreement between the AI and surgeons was slight, while inter-surgeon agreement was moderate (κ = 0.457–0.589). ROC analysis showed comparable AUC values (0.63–0.67) for all raters.

Discussion

The AI chatbot demonstrated diagnostic accuracy comparable to orthopedic trauma surgeons in directing treatment for isolated lateral malleolar fractures, despite using a systematically different measurement strategy. All raters exhibited conservative bias as comparted with the actual outcome with modest discriminatory ability, reflecting the inherent difficulty of this clinical issue. These findings support a potential complementary role for AI in ankle fracture triage, while final clinical management decisions should remain in the hands of the orthopedic surgeon.

Key words: lateral malleolar ankle fracture, artificial intelligence, surgical decision-making, radiographic assessment orthopedic surgery, diagnostic accuracy.

Acta Chir Orthop Traumatol Cech., 93, 2026, No. 3, p. 133–139

INTRODUCTION

Ankle fractures are among the most common injuries presenting to orthopedic emergency departments, with an estimated incidence of 169–179 per 100,000 person-years (12). Isolated lateral malleolar fracture at the level of the syndesmosis, constitutes a large subgroup and presents a particular management dilemma; this injury is essentially a mixed bag of bony and/or ligamentous injury ranging from potentially stable injuries that can be treated with a cast to combined unstable bony and ligamentous injury that can potentially displace. The diagnostic dilemma between a stable and unstable fracture has produced numerous studies (11, 21, 28). Missed instability risks such as malunion, post-traumatic arthrosis, and long-term functional impairment support operative treatment (5), while unnecessary surgery exposes patients to unnecessary risks (10).

Standardized radiographic parameters including, medial clear space (MCS), tibiofibular clear space (TFCS), and tibiofibular overlap (TFO), have been established as the cornerstones of ankle mortise stability assessment (1, 25). Further diagnostic studies such as stress radiography, weight-bearing imaging, and MRI have been proposed to reduce diagnostic uncertainty, but their routine use is constrained by availability, cost, and patient tolerance (1, 25).

Artificial intelligence (AI) has demonstrated promising diagnostic capabilities across multiple domains of medical imaging, including fracture detection, fracture classification, and treatment planning (13,15). More recently, large language models (LLMs) with integrated visual analysis capabilities have become commercially accessible, enabling direct radiographic interpretation without the need for purpose-built training datasets or dedicated hardware. However, strict validation against specialist clinicians using real-world outcomes as a reference standard has been limited, and existing studies have predominantly evaluated AI in clear pathological cases rather than the ambiguous presentations that most challenge clinical judgment (3, 29).

The aim of this study is to compare the diagnostic performance of a commercially available AI chatbot with that of three fellowship-trained orthopedic trauma surgeons in evaluating a purposefully enriched cohort of isolated 44-B1 ankle fractures, to characterize differences in radiographic measurements between AI and orthopedic surgeons and to compare final decisions for the need of operative treatment.

MATERIAL AND METHODS

Our study is a retrospective comparison study at an academic level I trauma center, comparing between the performance of a commercially available AI chatbot (Claude, Anthropic, San Francisco, CA, USA) and the decision-making process of three fellowship-trained orthopedic trauma surgeons in evaluating isolated lateral malleolar fractures. The study was reported in accordance with the STROBE guidelines for observational studies (30). Two board-certified orthopedic surgeons independently screened all 44-B1 lateral malleolus fractures presenting to the orthopedic emergency department between January 2024 and December 2025, to identify cases considered equivocal regarding operative versus non-operative management. The screening was based solely on the visual radiographic presentation, without performing formal measurements. Only cases in which both screening surgeons independently agreed that the treatment decision was uncertain were included, yielding a final group of 50 patients.

Eligible patients were adults (≥18 years) with AO/OTA 44B1 fracture who completed a set of 3 standard ankle radiographic views (anteroposterior, mortise, and lateral) of adequate diagnostic quality. Exclusion criteria were: prior surgical intervention on the ipsilateral tibia, fibula, or ankle; a history of malignancy or systemic bone disease affecting the lower extremity; an open or pathological fracture; polytrauma or concomitant ipsilateral fractures; significant degenerative joint disease obscuring radiographic landmarks; or an incomplete radiographic series.

Three trauma-trained orthopedic surgeons, blinded to one another's assessments and to the actual treatment decision, independently measured the following standardized ankle parameters on each case: medial clear space (MCS; normal ≤4 mm), tibiofibular clear space (TFCS; normal ≤6 mm), and tibiofibular overlap (TFO; normal ≥6 mm on AP, ≥1 mm on mortise), each on both the AP and mortise views (25, 26), on a standardized PACS system, (Fuji, etc.;). Using these measurements, each surgeon designated the case for operative or non-operative treatment. AI chatbot was given the same radiographic images using a standardized protocol (see Supplementary Material), in order to measure the same parameters on its own, and to give a treatment decision for each case. A research coordinator who was not part of the rating process collected the actual treatment decision for each patient from hospital medical records, which served as the reference standard.

Agreement on the surgical decision was assessed using Cohen's kappa for pairwise comparisons and Fleiss' kappa for multi-rater agreement. Continuous radiographic measurements were compared using paired t-tests, Wilcoxon signed-rank tests and intraclass correlation coefficients. Diagnostic accuracy against the real-life treatment decision was evaluated by sensitivity, specificity (Wilson score 95% confidence intervals), positive and negative predictive values, overall accuracy, and Cohen's kappa versus the reference standard. McNemar's test was used to detect systematic bias in each rater's recommendations relative to the actual outcome. A human majority consensus (agreement of ≥2 of 3 surgeons) was also derived as an additional comparator. Significance set at $p < 0.05$ (two-tailed).

Table 1. **Descriptive statistics of radiographic measurements and surgery rates**

PARAMETER	AI (CLAUDE)	SURGEON 1	SURGEON 2	SURGEON 3
MCS AP	4.25 ± 1.03	3.79 ± 1.34	3.99 ± 1.33	3.98 ± 1.45
TFCS AP	6.06 ± 0.82	4.16 ± 1.60	4.21 ± 1.30	4.80 ± 1.46
TFO AP	4.87 ± 2.16	6.39 ± 1.95	6.20 ± 2.10	7.51 ± 2.33
MCS Mortise	4.24 ± 1.03	3.66 ± 1.76	3.42 ± 1.63	3.79 ± 1.71
TFC Mortise	6.05 ± 0.83	3.88 ± 1.37	3.61 ± 1.30	3.79 ± 1.74
TFO Mortise	4.58 ± 1.93	2.37 ± 1.15	3.94 ± 1.68	3.85 ± 1.99
Surgery rate	20/50 (40.0%)	23/50 (46.0%)	19/50 (38.0%)	17/50 (34.0%)

Values are presented as mean ± SD (mm) for continuous measurements or n/N (%) for surgery rates. TCA AP = talar tilt angle; N/A = not available.

RESULTS

A total of 50 patients with 44B1 fractures were included in the analysis of our study. In real life, 28 patients (56.0%) underwent surgical fixation and 22 patients (44.0%) were managed non-operatively. All four raters recommended surgery at lower rates than the actual operative rate; the AI chatbot recommended surgery in 20 cases (40.0%), Surgeon 1 in 23 cases (46.0%), Surgeon 2 in 19 cases (38.0%), and Surgeon 3 in 17 cases (34.0%) of 50 evaluable cases, compared to 28 cases (56.0%) in real life outcomes. The human majority consensus (agreement of ≥ 2 of 3 surgeons) recommended surgery in 19 cases (38.0%) (Table 1; Fig. 1). The medial clear space (MCS) emerged as the strongest predictor across all raters. For the human mean measurements, MCS both on AP and mortise views were the only parameters significantly correlated with the real-life outcome, with higher values associated with operative management.

Among the 28 patients who actually underwent surgery, Ortho 1 demonstrated the highest sensitivity (78.6%), missing only 6 operated cases, while the AI chatbot, Ortho 2, and Ortho 3 performed comparable sensitivity, correctly identifying 53.6–57.1% of operated cases and missing 11–13 patients each. Among the 22 patients managed non-operatively, all raters achieved high specificity (81.8–95.5%) (Fig. 2).

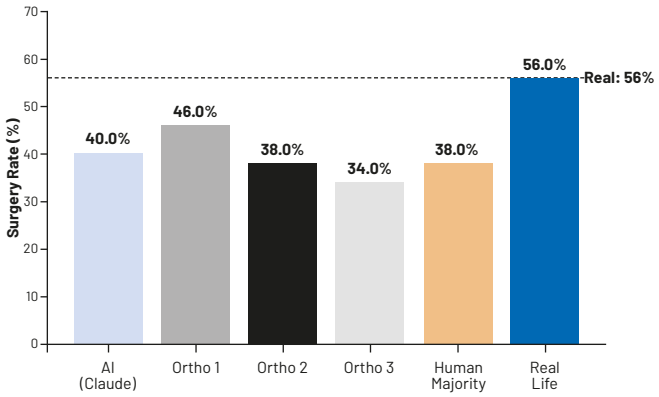

Fig. 1. **Surgery recommendation rate by rater versus real-life outcomes.**

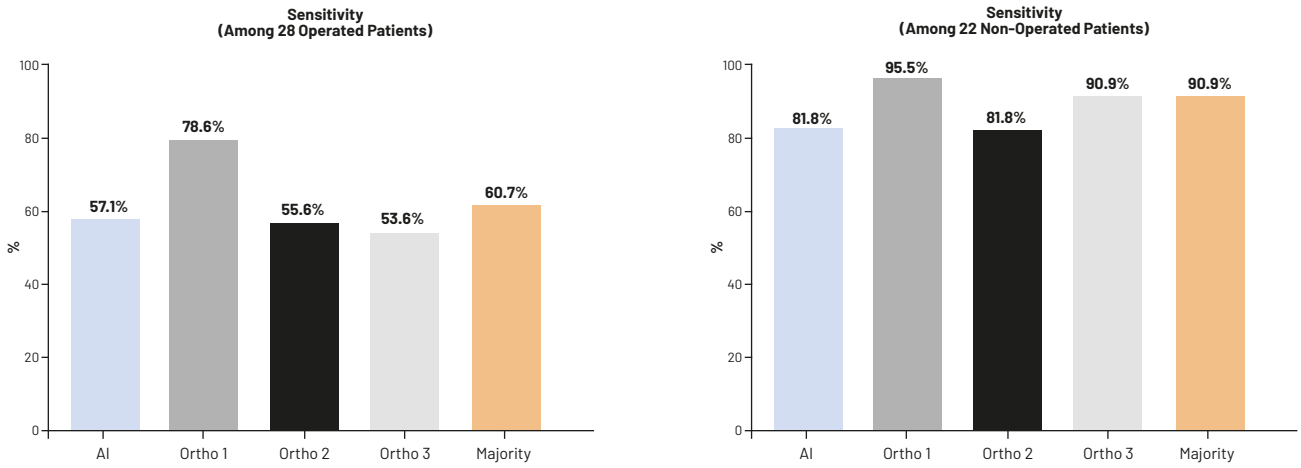

Fig. 2. **Subgroup analysis - sensitivity and specificity by rater (operated vs non-operated cases).**

Acta Chir Orthop Traumatol Cech., 93, 2026, No. 3, p. 133–139

Table 2. **Diagnostic accuracy of each rater against real-life surgical outcome**

RATER	N	TP	TN	FP	FN	SENS %	95% CI	SPEC %	95% CI	PPV %	NPV %	ACC %	P
AI (Claude)	50	16	18	4	12	57.1	(39.1–73.5)	81.8	(61.5–92.7)	80.0	60.0	68.0	0.08
Surgeon 1	50	22	21	1	6	78.6	(60.5–89.8)	95.5	(78.2–99.2)	95.7	77.8	86.0	0.13
Surgeon 2	50	15	18	4	12	55.6	(37.3–72.4)	81.8	(61.5–92.7)	78.9	60.0	67.3	0.08
Surgeon 3	50	15	20	2	13	53.6	(35.8–70.5)	90.9	(72.2–97.5)	88.2	60.6	70.0	0.01
Human Majority	50	17	20	2	11	60.7	(42.4–76.4)	90.9	(72.2–97.5)	89.5	64.5	74.0	0.03

TP = true positive; TN = true negative; FP = false positive; FN = false negative; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; PV = negative predictive value; Acc = accuracy; $p < 0.05$.

Among the 3 surgeons, the mean of calculated values (MCS, TFC, and TFO) was generally comparable for all parameters, with the greatest inter-surgeon variability observed in TFO mortise measurements (range 2.37–3.94 mm). Diagnostic accuracy metrics against the real-life treatment outcome are detailed in Table 2 and Figure 3. Surgeon 1 achieved the highest overall accuracy (86.0%), achieving substantial agreement with the real-life outcome ($\kappa = 0.723$). The AI achieved 68.0% accuracy, yielding fair agreement with reality ($\kappa = 0.37$). This was comparable to Surgeon 2 (accuracy 67.3%, $\kappa = 0.36$) and Surgeon 3 (accuracy 70.0%, $\kappa = 0.42$). The human majority consensus achieved 74.0% accuracy with moderate agreement ($\kappa = 0.49$). All raters demonstrated higher specificity than sensitivity, indicating a conservative bias, a tendency to under-recommend surgery relative to the actual treatment decisions. McNemar's test reached statistical significance for Surgeon 3 ($p = 0.01$) and the human majority ($p = 0.03$), confirming a systematic under-recommendation of surgery by these raters.

Inter-rater agreement between the AI and each individual surgeon was slight ($\kappa = 0.054$–0.146); and raw agreement of 55.1–58.0%. The inter-surgeon agreement was moderate, with pairwise kappa values ranging from 0.457 (Surgeons 1 vs. 2) to 0.589 (Surgeons 1 vs. 3) and raw agreement of 73.5–80.0%. McNemar's tests comparing the AI to each surgeon were non-significant (all $p > 0.05$), indicating that the disagreement was bi-directional rather than reflecting a systematic directional bias by the AI. The AI demonstrated significant systematic biases in five of six measurements (all $p < 0.005$), with only MCS AP showing no significant difference (mean difference +0.33 ± 1.46 mm, $p = 0.117$). The largest biases were observed for TFC mortise (+2.26 ± 1.30 mm), TFCS AP (+1.68 ± 1.36 mm), and TFO AP (–1.83 ± 2.81 mm). All ICC values between the AI and the human mean were poor (<0.21), indicating that the AI's measurements did not reliably reproduce those of the surgeons. Bland–Altman analysis confirmed wide 95% limits of agreement across all parameters, with the broadest range for TFO AP (–7.34 to +3.68 mm). The strongest inter-surgeon

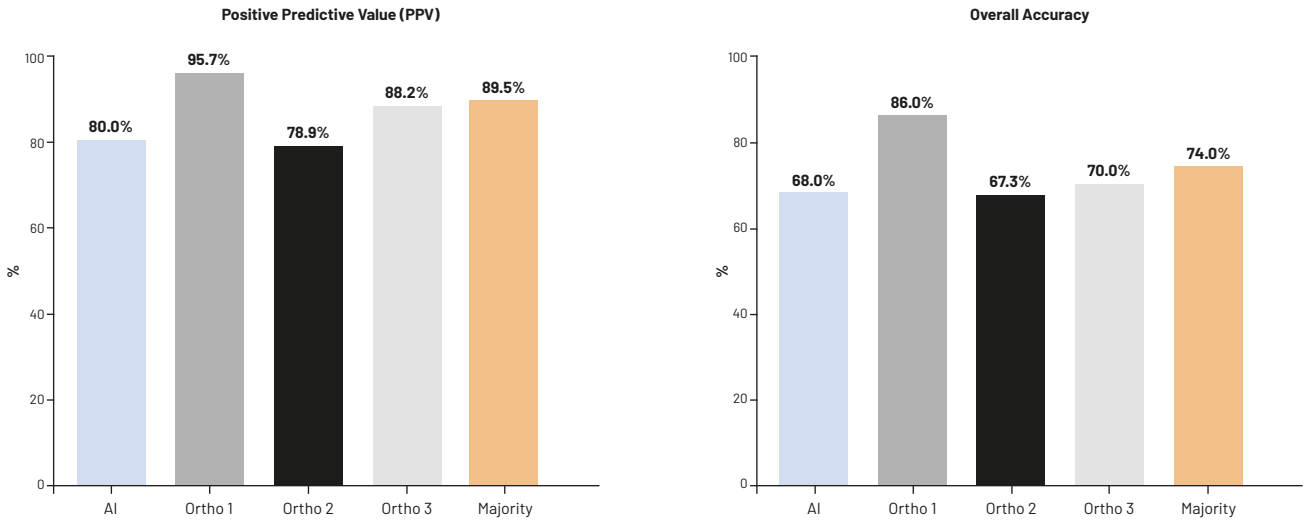

Fig. 3. **Diagnostic performance metrics: positive predictive value and overall accurancy.**

reliability was observed for MCS mortise (ICC 0.683–0.862, moderate to good) and MCS AP (ICC 0.686–0.751, moderate). TFO AP showed moderate agreement (ICC 0.599–0.622), while TFCS and TFC mortise showed poor-to-moderate reliability (ICC 0.375–0.566).

ROC analysis using a composite radiographic instability score (sum of standardized MCS, TFCS, and TFC values minus TFO values) is presented in the *supplementary materials section*. The area under the curve (AUC) was comparable across all raters: Ortho 1 achieved the highest AUC (0.667), followed by the AI (0.653), Ortho 3 (0.656), Ortho 2 (0.639), and the human mean (0.638). All AUC values fell in the 0.63–0.67 range, indicating modest discriminatory ability for all raters, with no single rater's measurements clearly superior in predicting the actual surgical outcome when combined into a composite score.

DISCUSSION

This study compared the diagnostic performance of a commercially available AI chatbot (Claude, Anthropic) with that of three trauma-trained orthopedic surgeons in evaluating borderline isolated lateral malleolar ankle fractures; cases in which experienced clinicians were themselves uncertain about the surgical indication (11, 22, 27). To our knowledge, this represents one of the first radiographic studies of a LLM with image analysis against a panel of specialty surgeons using real-life treatment decisions as the standard. Both the surgeons and the AI systematically under-recommended surgery relative to the actual operative rate, revealing a universal conservative bias in this clinically uncertain population. Although similar treatment decisions were made in some of the cases, the AI arrived to these decisions through a systematically different measurement strategy, overestimating tibiofibular clear space and underestimating tibiofibular overlap, two standard but imperfect radiographic surrogates of syndesmotic injury with known variability (5, 16, 20, 23).

The AI's overall accuracy was close to that of Surgeon 2 and Surgeon 3, but slightly lower than the consensus of the orthopedic surgeons. Surgeon 1 performed best, with much higher sensitivity (78.6% compared to 53.6–57.1% for the others) and the highest specificity (95.5%). The differences between surgeons are important, as many studies show that agreement on radiological findings can be low in ankle fractures and also in other fractures as well (1, 14, 18, 32). In our study, another orthopedic surgeon made independent decisions to reflect real-life practice, so it can be considered a 4th surgeon who provides his opinion on the radiographs. Comparing these results to the AI is interesting because they align with previous research showing significant differences among human observers in orthopedic radiographic assessments. This suggests that AI performance in fracture cases is not necessarily inferior to human expertise, but generally falls within the lower to middle range of human performance. Overall, these findings indicate that while the AI demonstrated a capacity to assess ankle mortise radiographs and render treatment decisions, its agreement with both human surgeons and real-life outcomes remained limited, performing comparably to two of the three surgeons but substantially below the best-performing surgeon. Although this contrasts with some recent reports of AI meeting or exceeding specialist-level accuracy in fracture detection, this difference may be explained by the greater complexity and clinical ambiguity of the ankle fracture cases in our cohort (2, 13, 17).

An interesting feature of our study is the universal conservative bias: the AI recommended surgery in only 40.0% of cases, and the three surgeons in 34.0–46.0%, against an actual operative rate of 56.0%. The tendency to under-recommend surgery in equivocal isolated lateral malleolar fractures has previously been reported in studies evaluating the Ottawa ankle rules and radiographic threshold criteria, and likely reflects a well-known conservatism in clinical decision-making when pathological thresholds are borderline (6, 19). In the AI, this bias may additionally reflect the inherent challenge of mapping continuous radiographic measurements to a binary operative decision when values cluster near established thresholds, a task that, as our ROC analysis demonstrates carries only modest discriminatory power for all raters regardless of human or AI origin. AI's false negative rate in this study (42.9%) therefore, represents the primary safety concern for any clinical deployment, and should be weighed carefully before considering AI-assisted triage in this setting. Notably, the AI and Surgeon 2 showed identical specificity (81.8%), suggesting that over-recommendation of surgery is not a predominant AI failure mode in this context. This finding reinforces that the borderline lateral malleolus fracture represents a difficult clinical decision in which radiographic measurements alone, whether obtained by experienced surgeons or AI, provide limited predictive power, and that clinical factors beyond standard radiographic parameters likely influence the ultimate treatment decision.

Despite arriving at surgical recommendations comparable to Surgeons 2 and 3, the AI's radiographic measurements were systematically different from all human raters across five of six parameters, with particularly large biases in tibiofibular clear space. The AI does not appear to replicate human measurement techniques but may instead extract decision-relevant signals through a different internal representation of radiographic instability. This is consistent with the tendency of large language models to solve tasks via emergent pattern recognition and internal feature representations rather than explicit rule application (24, 31). The MCS, which showed no significant AI bias and the highest intersurgeon reliability, emerged as the parameter most consistently shared between human and AI assessments and was the strongest predictor

of operative outcome across all raters. This convergence on MCS as the most reliable and clinically informative parameter corroborates its established primacy in ankle mortise assessment, where MCS widening is a key marker of deltoid insufficiency and ankle instability on stress and injury radiographs (4, 9).

Several limitations of this study deserve mention. First, the retrospective design and singlecenter cohort limit generalizability. Second, the reference standard was reallife treatment rather than a biomechanical ground truth such as stress radiography or intraoperative syndesmotic assessment, and true instability may not perfectly correlate with the operative decision, which can be influenced by patient factors, surgeon preference, and logistical considerations beyond the radiograph alone (7, 8). Third, the AI was evaluated using a single commercially available chatbot accessed through a standardized prompt protocol (see Supplementary Material); different AI architectures, prompt strategies, or purposebuilt medical AI systems may perform differently, and our findings may therefore not be generalizable to other AI implementations.

CONCLUSIONS

All raters showed a conservative bias with modest radiographic discriminatory ability, reflecting the difficulty of this clinical scenario rather than a specific AI limitation. The AI's different measurement strategy, yet similar treatment recommendations, suggests it uses distinct pattern-recognition pathways instead of reproducing human techniques. Overall, these findings support a complementary role for AI in ankle fracture triage, while final management decisions should remain with the orthopedic surgeon. ■

Supplementary material: AI chatbot protocol

Hello chat

You are an orthopedic surgeon, and you need to decide whether the following ankle fractures require surgery or treatment with a cast.

Please answer the following questions:

1. Calculate the following measurements:

MCS AP
TFCS AP
TFO AP
MCS Mortise
TFC Mortise
TFO Mortise

2. Does this patient needs a surgery?
   ROC Curves Comparing Discriminatory Ability of Raters Using Composite Radiographic Instability Score

References

1. Croft S, Furey A, Stone C, Moores C, Wilson R. Radiographic evaluation of the ankle syndesmosis. Can J Surg. 2015;58:58. doi:10.1503/cjs.004214.
2. Elbahi MK, Muhammed A, Mohamednour MFA, Mukhtar FS. Artificial intelligence in fracture diagnosis on radiographs: evidence, pitfalls, and pathways for clinical integration (2020–2025). Cureus. 2025;17:e93124. doi:10.7759/cureus.93124.
3. Erginoğlu SE, Ülgen NK, Yiğit N, Nazlıgül AS, Akkurt MO. Multimodal large language model for fracture detection in emergency orthopedic trauma: a diagnostic accuracy study. Diagnostics. 2026;16:476. doi:10.3390/diagnostics16030476.
4. Gibson PD, Ippolito JA, Hwang JS, Didesch J, Koury KL, Reilly MC, Adams M, Sirkin M. Physiologic widening of the medial clear space: what's normal? J Clin Orthop Trauma. 2019;10(Suppl 1):S62. doi:10.1016/j.jcot.2019.04.016.
5. Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti GM, Mangiavini L. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg. 2023;10:1284015. doi:10.3389/fsurg.2023.1284015.
6. Gomes YE, Chau M, Banwell HA, Causby RS. Diagnostic accuracy of the Ottawa ankle rule to exclude fractures in acute ankle injuries in adults: a systematic review and meta-analysis. BMC Musculoskelet Disord. 2022;23:885. doi:10.1186/s12891-022-05831-7.
7. Goodman AD, Blood TD, Benavent KA, Earp BE, Akelman E, Blazar PE. Implicit and explicit factors that influence surgeons' decision-making for distal radius fractures in older patients. J Hand Surg Am. 2022;47:719–726. doi: 10.1016/j.jhsa.2022.03.013.
8. Gunaratnam C, Bernstein M. Factors affecting surgical decision-making: a qualitative study. Rambam Maimonides Med J. 2018;9:e0003. doi:10.5041/rmmj.10324.
9. Hecht V, Mosimann ES, Krause F, Kurze C, Lustenberger T, Anwander H. The medial clearspace is a risk factor for secondary dislocation following cast immobilization after closed reduction in closed ankle fracture dislocations. Eur J Trauma Emerg Surg. 2025;51:161. doi:10.1007/s00068-025-02803-z.
10. Herrera-Pérez M, Valderrabano V, Godoy-Santos AL, de César Netto C, González-Martín D, Tejero S. Ankle osteoarthritis: comprehensive review and treatment algorithm proposal. EFORT Open Rev. 2022;7:448–459. doi:10.1530/EOR-21-0117.
11. Julian TH, Broadbent RH, Ward AE. Surgical vs non-surgical management of Weber B fractures: a systematic review. Foot Ankle Surg. 2020;26:494–502. doi:10.1016/j.fas.2019.06.006.

12. Juto H, Nilsson H, Morberg P. Epidemiology of adult ankle fractures: 1756 cases identified in Norrbotten County during 2009–2013 and classified according to AO/OTA. BMC Musculoskelet Disord. 2018;19:441. doi:10.1186/s12891-018-2326-x.
13. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, Stewart M, Collins GS, Furniss D. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. 2022;304:50–62. doi: 10.1148/radiol.211785.
14. Lakomkin N, Fabricant PD, Cruz AI, Brusalis CM, Chauvin NA, Todd J. Interrater reliability and age-based normative values for radiographic indices of the ankle syndesmosis in children. JBJS Open Access. 2016;2:e0004. doi:10.2106/JBJS.OA.16.00004.
15. Langerhuizen DWG, Janssen SJ, Mallee WH, van den Bekerom MPJ, Ring D, Kerkhoffs GMMJ, Jaarsma RL, Doornberg JN. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. Clin Orthop Relat Res. 2019;477:2482–2491. doi: 10.1097/CORR.0000000000000848.
16. Mergen M, Spitzl D, Ketzer C, Strenzke M, Marka AW, Makowski MR, Bressem KK, Adams LC, Gassert FT. Leveraging large language models for accurate AO fracture classification from CT. text reports. J Imaging Informatics Med. 2026;39:1861–1867. doi:10.1007/s10278-025-01603-6.
17. Nowroozi A, Salehi MA, Shobeiri P, Agahi S, Momtazmanesh S, Kaviani P, Kalra MK. Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: a systematic review and meta-analysis. Clin Radiol. 2024;79:579–588. doi:10.1016/j.crad.2024.04.009.
18. Oulianski M, Avraham D, Lubovsky O. Radiographic evaluation of distal radius fracture healing by time: orthopedist versus qualitative assessment of image processing. Trauma Care. 2022;2:481–486. doi:10.3390/traumacare2030040.
19. Pires R, Pereira A, Abreu-e-Silva G, Labronici P, Figueiredo L, Godoy-Santos A, Kfuri M. Ottawa ankle rules and subjective surgeon perception to evaluate radiograph necessity following foot and ankle sprain. Ann Med Health Sci Res. 2014;4:432. doi:10.4103/2141-9248.133473.
20. Pogliacomi F, De Filippo M, Casalini D, Longhi A, Tacci F, Perotta R, Pagnini F, Tocco S, Ceccarelli F. Acute syndesmotic injuries in ankle fractures: from diagnosis to treatment and current concepts. World J Orthop. 2021;12:270. doi:10.5312/wjo.v12.i5.270.
21. Reyes-Valdés A, Martínez-Ledezma M, Fernández-Quezada D, Guzmán-Esquivel J, Cárdenas-Rojas MI. Prevalence and characteristics of patients requiring surgical reinterventions for ankle fractures. J Clin Med. 2023;12:5843. doi:10.3390/jcm12185843.
22. Rooney EM, Finney FT, Talusan P, Holmes JR, Walton D. Mid term 5-year follow up of a novel algorithm for non-operative Weber B ankle fractures. Foot Ankle Orthop. 2019;4:2473011419S00366. doi:10.1177/2473011419s00366.
23. Smith AM, Jacquez EA, Argintar EH. Assessing the efficacy of an AI-powered chatbot (ChatGPT) in providing information on orthopedic surgeries: a comparative study with expert opinion. Cureus. 2024;16:e63287. doi:10.7759/cureus.63287.
24. Sorin V, Klang E. Large language models and the emergence phenomena. Eur J Radiol Open. 2023;10:100494. doi:10.1016/j.ejro.2023.100494.
25. Strash WW, Berardo P. Radiographic assessment of the hindfoot and ankle. Clin Podiatr Med Surg. 2004;21:295–304. doi:10.1016/j.cpm.2004.03.004.
26. Surmanowicz P, Hamilton AM, Mondal P, Kulyk P, Sahota N, Obaid H. Correlation between Weber classification of ankle fractures and medial clear space widening on radiography. Diagnostics (Basel). 2025;15:15162085. doi:10.3390/diagnostics15162085.
27. Tansey PJ, Chen J, Panchbhavi VK. Current concepts in ankle fractures. J Clin Orthop Trauma. 2023;45:102260. doi:10.1016/j.jcot.2023.102260.
28. Toru HK, Khan AA, Ali N. Operative vs. nonoperative management of isolated Weber B ankle fractures. Cureus. 2025;17:e78028. doi:10.7759/cureus.78028.
29. von der Stück MS, Vuskov R, Westfechtel S, Siepmann R, Kuhl C, Truhn D, Nebelung S. Visual large language models in radiology: a systematic multimodel evaluation of diagnostic accuracy and hallucinations. Life. 2026;16(1):66. doi:10.3390/life16010066.
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453-1457.
31. Wei J, Tay Y, Bommasani R, E Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W. Emergent abilities of large language models. TMLR. 2022; 2206. http://arXiv.org/abs/2206.07682.
32. Whelan DB, Bhandari M, McKee MD, Guyatt GH, Kreder HJ, Stephen D, Schemitsch EH. Interobserver and intraobserver variation in the assessment of the healing of tibial fractures after intramedullary fixation. J Bone Joint Surg Br. 2002;84:15–18. doi:10.1302/0301-620x.84b1.11347.