Evaluation of Artificial Intelligence (ChatGPT-5.2) in the Classification and Indication for Fixation of Posterior Malleolar Fractures: A Multicenter External Validation Study

Héctor A. Rivadeneira Jurado,[*] Elías A. Rivadeneira Jurado,[*] Daniel Espinoza Freire,[*] Andrés F. Samaniego,[*] Ezequiel Lulkin,[*] Sebastián Pereira,[*] Fernando Bidolegui,[**] Tomás Macagno[**]

[*]*Orthopedics and Traumatology Service, Hospital Sirio-Libanés, Autonomous City of Buenos Aires, Argentina*

[**]*Orthopedics and Traumatology Service, Sanatorio Otamendi y Miroli, Autonomous City of Buenos Aires, Argentina*

ABSTRACT

Introduction: Posterior malleolar fractures have a significant impact on ankle joint congruity. The indication for fixation no longer depends solely on fragment size but also on fracture morphology. Artificial intelligence (AI) has emerged as a tool to support clinical decision-making. The objective of this study was to evaluate the ability of AI to classify posterior malleolar fractures and determine the indication for fixation, compared with a reference standard based on expert consensus. **Materials and Methods:** A retrospective diagnostic accuracy study with external validation was conducted in accordance with the STARD-AI and GAMER guidelines. A protocol based on the Bartoníček and Rammelt classification was developed using 24 cases for calibration. Subsequently, 9 cases were evaluated using radiographs and computed tomography scans and analyzed by 12 experts and the ChatGPT-5.2 model. Agreement in fracture classification and sensitivity for the indication for fixation were assessed using Cohen's kappa coefficient. **Results:** ChatGPT-5.2 achieved 78% agreement in fracture classification, with a kappa coefficient of 0.56, indicating moderate agreement. Sensitivity for the indication for posterior malleolar fixation was 100%. **Conclusions:** Artificial intelligence demonstrated performance comparable to that of experts in the classification of posterior malleolar fractures and high sensitivity in determining the indication for fixation. It proved useful as a supportive tool in medical education settings. Studies with larger sample sizes are needed to validate these findings.

Keywords: Artificial intelligence; posterior malleolus; multicenter study.

Level of Evidence: III

Evaluación de la capacidad de la inteligencia artificial (ChatGPT-5.2) para clasificar fracturas del maléolo posterior e indicar su fijación: estudio multicéntrico de validación externa

RESUMEN

Introducción: Las fracturas del maléolo posterior del tobillo tienen un gran impacto en la congruencia articular del tobillo. La indicación de fijación ya no depende exclusivamente del tamaño del fragmento, sino también de su morfología. La inteligencia artificial surge como una herramienta para apoyar la toma de decisiones clínicas. El objetivo de este estudio fue evaluar la capacidad de la inteligencia artificial para clasificar fracturas del maléolo posterior e indicar su fijación, comparada con la de un estándar de referencia basado en el consenso de expertos. **Materiales y Métodos:** Se realizó un estudio retrospectivo de exactitud diagnóstica con validación externa, siguiendo las guías STARD-AI y GAMER. Se diseñó un protocolo basado en la clasificación de Bartoníček y Rammelt, utilizando 24 casos para calibración. Se evaluaron 9 casos mediante radiografías y tomografía computarizada, analizados por 12 expertos y por el modelo ChatGPT-5.2. Se determinó la concordancia en la clasificación y la sensibilidad para la indicación de fijación, utilizando el coeficiente kappa de Cohen. **Resultados:** El ChatGPT-5.2 alcanzó una concordancia del 78% en la clasificación de fracturas, con un coeficiente kappa de 0,56, que indica una concordancia moderada. La sensibilidad para la

Received on April 22nd, 2026. Accepted after evaluation on May 10th, 2026 • Dr. HÉCTOR A. RIVADENEIRA JURADO • 1bhribadeneirajurado@gmail.com https://orcid.org/0009-0008-6397-9718

How to cite this article: Rivadeneira Jurado HA, Rivadeneira Jurado EA, Espinoza Freire D, Samaniego AF, Lulkin E, Pereira S, et al. Evaluation of Artificial Intelligence (ChatGPT-5.2) in the Classification and Indication for Fixation of Posterior Malleolar Fractures: A Multicenter External Validation Study. *Rev Asoc Argent Ortop Traumatol* 2026;91(3):246-249. https://doi.org/10.15417/issn.1852-7434.2026.91.3.2348

indicación de fijación del maléolo posterior fue del 100%. **Conclusiones:** La inteligencia artificial tuvo un desempeño comparable al de los expertos en la clasificación de fracturas del maléolo posterior y una alta sensibilidad en la indicación de fijación. Resultó útil como herramienta de apoyo en contextos de formación médica. Se requieren estudios con muestras más grandes para validar estos hallazgos.
Palabras clave: Inteligencia artificial; maléolo posterior; estudio multicéntrico.
Nivel de Evidencia: III

INTRODUCTION

Posterior malleolar fractures have assumed a central role in the contemporary management of ankle fractures, not only because of their frequency but also because of their direct impact on syndesmotic stability and tibiotalar joint congruity. Current evidence has moved beyond the traditional paradigm based exclusively on fragment size and indicates that variables such as fracture morphology, involvement of the fibular incisura, and the degree of articular displacement are key determinants of the indication for fixation and the patient's functional outcome.[1,2]

In this context, the systematic use of computed tomography has enabled more accurate characterization of these injuries. The Bartoníček and Rammelt classification has been shown to be clinically useful because it integrates the morphology of the posterior fragment with its biomechanical relevance, thereby facilitating individualized surgical decision-making.[3] However, interpretation of these imaging studies continues to depend on the surgeon's experience, and interobserver variability persists, even among specialists.

At the same time, the development of artificial intelligence (AI) models has emerged as a promising tool in orthopedics, particularly for fracture detection and classification using imaging studies. Recent research has shown that these systems can achieve levels of accuracy comparable to those of expert clinicians in certain settings and may also improve diagnostic performance when used as decision-support tools.[4-6] Nevertheless, their application to specific surgical decision-making, such as determining the indication for posterior malleolar fixation, remains limited and has been scarcely validated in the current medical literature.

In this context, the aim of this study was to evaluate the ability of an AI model to classify posterior malleolar fractures according to the Bartoníček and Rammelt classification and to determine the indication for fixation, using an expert consensus reference standard for comparison.

MATERIALS AND METHODS

A retrospective diagnostic accuracy study with external validation was conducted in accordance with the STARD-AI (Standards for Reporting Diagnostic Accuracy–Artificial Intelligence) and GAMER guidelines.

The study was carried out in two phases. In the first phase, a prompt was developed based on anatomical information and the Bartoníček and Rammelt classification to create a standardized evaluation protocol. Ninety-five ankle fracture cases were initially selected; 45 were reviewed, and 24 met the inclusion criteria and were used to calibrate the protocol before external validation. In addition, 9 cases were selected and sent to 12 independent volunteer experts, who were asked to classify each fracture according to the Bartoníček and Rammelt classification and determine whether posterior malleolar fixation was indicated. Each case included anteroposterior, mortise, and lateral ankle radiographs together with computed tomography (CT) images in the axial and sagittal planes (Figure). Data collection was performed using questionnaires created in Google Forms®.

In the second phase, the interpretations of the 12 experts and those generated by ChatGPT-5.2 acting as an expert evaluator were compared with the reference standard previously established from the patients' medical records.

The inclusion criteria were patients with ankle fractures involving the posterior malleolus, complete imaging studies (anteroposterior, mortise, and lateral radiographs together with CT scans), and complete medical records from admission through postoperative follow-up. The exclusion criteria were distal tibial fractures with secondary extension into the posterior malleolus and absence of postoperative follow-up. Fracture classification and the indication for posterior malleolar fixation were analyzed. Agreement was expressed as percentages and assessed using Cohen's kappa coefficient.

Figure. Sequence of images presented to ChatGPT-5.2 for interpretation. Ankle radiographs: anteroposterior (**A**), mortise (**B**), and lateral (**C**) views; and computed tomography images: axial and sagittal slices (**D and E**).

RESULTS

ChatGPT-5.2 achieved 78% agreement with the expert consensus reference standard for the classification of posterior malleolar fractures. The estimated Cohen's kappa coefficient was approximately 0.56, indicating moderate agreement. Regarding the indication for posterior malleolar fixation, ChatGPT-5.2 achieved a sensitivity of 100%, correctly identifying all cases in which fixation was indicated; no false-negative results were observed in the study cohort. The analyzed parameters are summarized in the Table.

Table. Diagnostic performance of the ChatGPT-5.2 model.

Parameter	Result
Total number of cases	9
Classification agreement	78%
Kappa coefficient (estimated)	0.56
Sensitivity for fixation	100%
False negatives	0

It is noteworthy that ChatGPT-5.2 demonstrated higher accuracy in posterior malleolar fracture patterns with greater displacement, whereas discrepancies were observed in fracture patterns with minimal displacement.

DISCUSSION

The results of this study demonstrate that AI can achieve levels of agreement comparable to those of experts in the evaluation of posterior malleolar fractures, particularly with respect to the indication for fixation.

The 100% sensitivity observed is clinically relevant, as failure to fix the posterior malleolus may be associated with persistent instability and poor functional outcomes.[1,2]

These findings are consistent with those of recent studies demonstrating the potential of AI for fracture diagnosis. Rivadeneira et al. reported perfect agreement between AI and expert evaluators in the classification of complex fractures.[7]

Similarly, Husarek et al., in a systematic review and meta-analysis, found that the use of AI as a decision-support tool increased diagnostic sensitivity, particularly among less experienced readers, compared with unassisted interpretation.[8]

Conversely, Mohammadi et al. reported that experts achieved higher diagnostic sensitivity than AI models, such as ChatGPT-4, when interpreting knee radiographs, suggesting that AI performance may still be inferior in certain clinical scenarios.[9]

Our study has important limitations. The small sample size limits the generalizability of the findings. In addition, the AI model was evaluated in a controlled environment, which may not fully reflect real-world clinical practice. Therefore, studies including larger samples and external validation are warranted.

Despite these limitations, the use of methodological guidelines such as STARD-AI and GAMER strengthens the validity of the study by enhancing transparency, standardization, and reproducibility in research on AI applications in orthopedic trauma.

CONCLUSIONS

ChatGPT-5.2 achieved 78% agreement with the reference standard and a Cohen's kappa coefficient of 0.56, indicating moderate agreement, while demonstrating high sensitivity for identifying cases requiring posterior malleolar fixation. It may serve as a useful decision-support tool in training settings for inexperienced physicians.

Conflicts of interest: The authors declare no conflicts of interest.

E. A. Rivadeneira Jurado ORCID ID: https://orcid.org/0009-0006-5784-5700
D. Espinoza Freire ORCID ID: https://orcid.org/0009-0000-9882-6027
A. F. Samaniego ORCID ID: https://orcid.org/0000-0002-6616-6471
E. Lulkin ORCID ID: https://orcid.org/0000-0002-4119-0483
S. Pereira ORCID ID: https://orcid.org/0000-0001-9475-3158
F. Bidolegui ORCID ID: https://orcid.org/0000-0002-0502-2300
T. Macagno ORCID ID: https://orcid.org/0009-0006-5009-9944

REFERENCES

1. Terstegen J, Weel H, Frosch KH, Rolvien T, Schlickewei C, Mueller E. Classifications of posterior malleolar fractures: a systematic literature review. *Arch Orthop Trauma Surg* 2023;143(7):4181-220. https://doi.org/10.1007/s00402-022-04643-7
2. Mohamed A, Fuad U, Elasad A, Shrestha S, Hagroo A, Pengas IP. Posterior malleolar fractures: From the „Forgotten Fragment" to modern concepts in management. *Cureus* 2025;17(10):e94681. https://doi.org/10.7759/cureus.94681
3. Bartoníček J, Rammelt S, Tuček M, Naňka O. Posterior malleolar fractures of the ankle. *Eur J Trauma Emerg Surg* 2015;41(6):587-600. https://doi.org/10.1007/s00068-015-0560-6
4. Verhage SM, Hoogendoorn JM, Krijnen P. When and how to operate the posterior malleolus fragment in trimalleolar fractures. *Arch Orthop Trauma Surg* 2018;138(9):1213-22. https://doi.org/10.1007/s00402-018-2949-2
5. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. Preprint. *Digit Med* 2017. https://doi.org/10.48550/arXiv.1711.06504
6. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci USA* 2018;115(45):11591-6. https://doi.org/10.1073/pnas.1806905115
7. Rivadeneira Jurado HA, Rivadeneira Jurado EA, Espinoza Freire D, Samaniego AF, Lulkin E, Bidolegui F, et al. Evaluación de la clasificación de las fracturas de platillo tibial según Schatzker-Kfuri utilizando radiografías y tomografía. Comparación entre el observador experto y el modelo ChatGPT-4o. *Rev Asoc Argent Ortop Traumatol* 2025;90(6):556-60. https://doi.org/10.15417/issn.1852-7434.2025.90.6.2224
8. Husarek J, Hess S, Razaeian S, Ruder TD, Sehmisch S, Müller M, et al. Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. *Sci Rep* 2024;14(1):23053. https://doi.org/10.1038/s41598-024-73058-8
9. Mohammadi S, Parviz S, Parvaz P, Pirmoradi MM, Afzalimoghaddam M, Mirfazaelian H. Diagnostic performance of ChatGPT in tibial plateau fracture in knee X-ray. *Emerg Radiol* 2025;32(1):59-64. https://doi.org/10.1007/s10140-024-02298-y.