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

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

[*]*Servicio de Ortopedia y Traumatología, Hospital Sirio-Libanés, Ciudad Autónoma de Buenos Aires, Argentina*

[**]*Servicio de Ortopedia y Traumatología, Sanatorio Otamendi y Miroli, Ciudad Autónoma de Buenos Aires, Argentina*

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 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

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

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

Recibido el 22-4-2026. Aceptado luego de la evaluación el 10-5-2026 • Dr. HÉCTOR A. RIVADENEIRA JURADO • 1bhribadeneirajurado@gmail.com https://orcid.org/0009-0008-6397-9718

Cómo citar este artículo: Rivadeneira Jurado HA, Rivadeneira Jurado EA, Espinoza Freire D, Samaniego AF, Lulkin E, Pereira S, et al. 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. *Rev Asoc Argent Ortop Traumatol* 2026;91(3):246-249. https://doi.org/10.15417/issn.1852-7434.2026.91.3.2348

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

INTRODUCCIÓN

Las fracturas del maléolo posterior han cobrado un rol protagónico en el manejo contemporáneo de las fracturas de tobillo, no solo por su frecuencia, sino también por su impacto directo en la estabilidad sindesmótica y la congruencia de la articulación tibioastragalina. La evidencia actual ha desplazado el paradigma clásico basado exclusivamente en el tamaño del fragmento, y sostiene que variables, como la morfología del trazo, el compromiso de la incisura peronea y el grado de desplazamiento articular, constituyen factores determinantes en la indicación de fijación y en el pronóstico funcional del paciente.[1,2]

En este contexto, la incorporación sistemática de la tomografía computarizada ha permitido caracterizar estas lesiones con más precisión. Se ha demostrado que la clasificación propuesta por Bartoníček y Rammelt es clínicamente útil al integrar la morfología del fragmento posterior con su relevancia biomecánica, facilitando la toma de decisiones quirúrgicas individualizadas.[3] Sin embargo, la interpretación de estos estudios continúa dependiendo de la experiencia del cirujano, y hay variabilidad interobservador, incluso entre especialistas.

Al mismo tiempo, el desarrollo de modelos de inteligencia artificial (IA) ha emergido como una herramienta prometedora en el campo de la traumatología, particularmente en la detección y clasificación de fracturas mediante estudios por imágenes. Según investigaciones recientes, estos sistemas pueden alcanzar niveles de precisión comparables con los de expertos en determinados escenarios, y también pueden mejorar el rendimiento diagnóstico cuando se utilizan como herramientas de apoyo.[4-6] No obstante, su aplicación en la toma de decisiones quirúrgicas específicas, como la indicación de fijación del maléolo posterior, sigue siendo limitada y escasamente validada en la literatura médica actual.

En este escenario, el objetivo de este estudio fue evaluar la capacidad de un modelo de IA para clasificar fracturas del maléolo posterior según la clasificación de Bartoníček y Rammelt, e indicar su fijación, comparada con la de un estándar de referencia basado en el consenso de expertos.

MATERIALES Y MÉTODOS

Se llevó a cabo un estudio retrospectivo de exactitud diagnóstica con validación externa, siguiendo las guías STARD-AI (*Standards for Reporting Diagnostic Accuracy – Artificial Intelligence*) y GAMER.

El estudio se realizó en dos fases: la primera se creó a través de un *prompt* que se estructuró con información de anatomía, la clasificación de Bartoníček y Rammelt, para crear un protocolo para el cual se seleccionaron 95 casos de fracturas de tobillo, 45 de ellos fueron evaluados, 24 cumplieron los criterios de inclusión, y se usaron para calibrar el protocolo antes de la validación externa. Asimismo, se seleccionaron 9 casos que fueron enviados a 12 expertos independientes y voluntarios, para analizar cada caso clasificando la fractura según Bartoníček y Rammelt e indicando la fijación o no del maléolo posterior. Cada uno de los casos contaba con radiografías de tobillo, en proyecciones anteroposterior, de mortaja y de perfil, y una tomografía computarizada con cortes axial y de perfil (Figura). La recopilación del análisis se obtuvo mediante encuestas creadas en Google Forms®.

En la segunda fase del estudio, se realizó el análisis de interpretación de los 12 expertos, el ChatGPT-5.2 como experto, y el resultado con el estándar de referencia definido previamente con la información de las historias clínicas.

Por otro lado, los criterios de inclusión fueron: pacientes con una fractura de tobillo con compromiso del maléolo posterior, estudios completos: radiografías anteroposterior, de mortaja y de perfil, y tomografía computarizada; e historia clínica completa desde el ingreso hasta el control posoperatorio. Los criterios de exclusión fueron: pacientes con una fractura de tibia distal con extensión secundaria al maléolo posterior y falta de seguimiento posoperatorio. Se realizaron los análisis de clasificación de la fractura e indicación de fijación del maléolo posterior. El análisis se describió, en forma porcentual, con el coeficiente de correlación kappa de Cohen.

Figura. Secuencia de imágenes presentadas para la interpretación del ChatGPT-5.2. Radiografías de tobillo, anteroposterior (**A**), de mortaja (**B**), de perfil (**C**), y tomografía computarizada, cortes axial y de perfil (**D y E**).

RESULTADOS

El ChatGPT-5.2 alcanzó una concordancia del 78% respecto al estándar de referencia basado en el consenso de expertos al clasificar las fracturas del maléolo posterior. El coeficiente kappa estimado fue de aproximadamente 0,56, lo cual indica una concordancia moderada. Por otro lado, respecto a la indicación de fijación del maléolo posterior, el ChatGPT-5.2 tuvo una sensibilidad del 100%, identificó correctamente todos los casos en los que la fijación estaba indicada; no se registraron resultados falsos negativos en la cohorte analizada. En la Tabla, se resumen los parámetros analizados.

Tabla. Rendimiento diagnóstico del modelo ChatGPT-5.2

Parámetro	Resultado
Número total de casos	9
Concordancia en la clasificación	78%
Coeficiente kappa (estimado)	0,56
Sensibilidad para la fijación	100%
Falsos negativos	0

Cabe mencionar que el ChatGPT-5.2 tuvo una precisión más alta en patrones de fractura del maléolo posterior con mayor desplazamiento y que las discrepancias se observaron en casos con patrones de fractura sin gran desplazamiento.

DISCUSIÓN

Los resultados de este estudio demuestran que la IA puede alcanzar niveles de concordancia comparables a los de los expertos en la evaluación de fracturas del maléolo posterior, particularmente en la indicación de fijación.

La sensibilidad obtenida del 100% es clínicamente relevante, ya que omitir la fijación del maléolo posterior puede asociarse a inestabilidad persistente y malos resultados funcionales.[1,2]

Estos hallazgos coinciden con los de estudios recientes que han mostrado el potencial de la IA para diagnosticar fracturas. Rivadeneira y cols. señalan que la IA tiene una concordancia perfecta con los expertos al clasificar fracturas complejas.[7]

De manera similar, Husarek y cols., en una revisión sistemática y metanálisis, comprobaron que el uso de la IA como herramienta de apoyo incrementa la sensibilidad diagnóstica, especialmente en evaluadores con menos experiencia, en comparación con la interpretación sin asistencia.[8]

Por otro lado, Mohammadi y cols. comunicaron que la sensibilidad diagnóstica de los expertos al interpretar radiografías de rodilla fue más alta que la de los modelos de IA, como ChatGPT-4, esto refleja que el rendimiento de la IA aun puede ser inferior en determinados escenarios clínicos.[9]

Nuestro estudio tiene limitaciones importantes. El tamaño reducido de la muestra impide la generalización de los resultados. Además, el modelo de IA fue evaluado en un entorno controlado, lo que puede no reflejar completamente la práctica clínica real. En este sentido, se requieren estudios con una muestra más grande y validación externa.

A pesar de estas limitaciones, el empleo de guías metodológicas, como STARD-AI y GAMER, fortalece la validez del estudio, aportando transparencia, estandarización y reproducibilidad en la investigación de la IA aplicada a la traumatología.

CONCLUSIONES

La IA (ChatGPT-5.2) tuvo una concordancia del 78%, con un coeficiente kappa de 0,56, lo cual indica una concordancia moderada y alta sensibilidad para indicar la fijación del maléolo posterior. Es una herramienta de apoyo útil en escenarios de entrenamiento para médicos inexpertos.

Conflicto de intereses: Los autores no declaran conflictos de intereses.

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

BIBLIOGRAFÍA

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.