Revista Española de Cirugía Ortopédica y Traumatología xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Revista Española de Cirugía Ortopédica y Traumatología

journal homepage: www.elsevier.es/rot

Original

Fiabilidad de la inteligencia artificial (ChatGPT) en el diagnóstico y clasificación de las fracturas de meseta tibial

Reliability of artificial intelligence (ChatGPT) in the diagnosis and classification of tibial plateau fractures

C. Castillejo [a,*], M. Zapatero [a], J.M. Bogallo [a], F. Lorente [a], C. Ortiz [a], J. Romero [a], F. Rivas-Ruiz [b] y M.L. Bertrand [a]

[a] *Servicio de Cirugía Ortopédica y Traumatología, Hospital Universitario Costa del Sol, Universidad de Málaga, Marbella, Málaga, España*
[b] *Unidad de Apoyo a la Investigación, Hospital Universitario Costa del Sol, Universidad de Málaga, Marbella, Málaga, España*

INFORMACIÓN DEL ARTÍCULO

Palabras clave:
Fractura de meseta tibial
Inteligencia artificial
ChatGPT

RESUMEN

Introducción: Las fracturas de meseta tibial constituyen aproximadamente el 1% de todas las fracturas en adultos y el 8% en personas mayores de 65 años. Sin embargo, la tasa de error sigue siendo elevada debido a interpretaciones erróneas de las radiografías. La inteligencia artificial (IA) se presenta como una herramienta prometedora para mejorar la precisión diagnóstica, al permitir detectar y clasificar fracturas de forma automatizada, reduciendo tanto la variabilidad en la interpretación como la necesidad de pruebas complementarias.
Objetivo: Comparar la precisión en el diagnóstico y la clasificación de fracturas de meseta tibial mediante radiografías simples entre 3 grupos: facultativos especialistas de área, médicos internos residentes e inteligencia artificial (ChatGPT-4).
Métodos: Estudio observacional, descriptivo y transversal, realizado en una cohorte retrospectiva de pacientes atendidos entre 2020 y 2024 por fractura de meseta tibial. Se evaluaron radiografías anteroposteriores de forma ciega por 3 grupos: 3 facultativos especialistas de área, 3 médicos internos residentes y el modelo ChatGPT-4.0. Todos los evaluadores clasificaron las fracturas según la clasificación de Schatzker utilizando un mismo cuestionario estructurado para garantizar uniformidad en el proceso diagnóstico. El estándar de referencia fue el TAC. Para evaluar la concordancia entre evaluadores se empleó el índice Kappa para la detección de fractura y el Kappa ponderado por Ciccetti para la clasificación del grado de fractura, con intervalos de confianza del 95%. Se estableció un nivel de significación estadística de $p < 0,01$.
Resultados: Se incluyeron 387 radiografías, de las cuales 129 presentaban fracturas de meseta tibial (clasificadas según Schatzker: 7 tipo I, 28 tipo II, 5 tipo III, 16 tipo IV, 21 tipo V y 52 tipo VI) y 258 no presentaban fractura. La IA mostró la mayor precisión en la detección de fracturas, alcanzando un acuerdo absoluto del 99,5% y un índice Kappa de 0,98 (IC 95%: 0,97-1,00, $p < 0,001$), en comparación con el 97% obtenido por los facultativos ($K = 0,93$, IC 95%: 0,91-0,95, $p < 0,001$) y el 93% de los residentes ($K = 0,848$, IC 95%: 0,81-0,88, $p < 0,001$). En términos de variabilidad interobservador para el diagnóstico de fractura, la IA presentó mayor consistencia y menor variabilidad que los profesionales médicos. Sin embargo, en la clasificación del tipo de fractura, los adjuntos obtuvieron un índice Kappa ponderado superior (0,616, IC 95%: 0,554-0,679, $p < 0,001$) en comparación con la IA (0,612, IC 95%: 0,502-0,722, $p < 0,001$) y los residentes (0,572, IC 95%: 0,510-0,635, $p < 0,001$).
Conclusiones: La IA mostró una capacidad destacada en la detección de fracturas de meseta tibial, con niveles de precisión superiores a los observados en médicos residentes y adjuntos en este aspecto concreto. No obstante, en la clasificación según el sistema de Schatzker, los médicos adjuntos obtuvieron mejores resultados. Estos hallazgos sugieren que la IA puede constituir una herramienta de apoyo útil en el proceso diagnóstico, especialmente en etapas iniciales, complementando —pero no reemplazando— el juicio clínico y la experiencia del profesional de la salud.

Nivel de evidencia: nivel III. Diagnóstico. Estudio observacional descriptivo y transversal con grupo control.

* Autor para correspondencia.
Correo electrónico: coralcastillejoiniesta@gmail.com (C. Castillejo).

https://doi.org/10.1016/j.recot.2025.08.004
Recibido el 12 de marzo de 2025; Aceptado el 11 de agosto de 2025
On-line xxx

Cómo citar este artículo: C. Castillejo, M. Zapatero, J.M. Bogallo et al., Fiabilidad de la inteligencia artificial (ChatGPT) en el diagnóstico y clasificación de las fracturas de meseta tibial, Revista Española de Cirugía Ortopédica y Traumatología, https://doi.org/10.1016/j.recot.2025.08.004

Keywords:
Tibial plateau fracture
Artificial Intelligence
ChatGPT

ABSTRACT

Objective: To compare the diagnostic and classification accuracy of tibial plateau fractures on simple radiographs among three groups: knee surgeons, resident physicians, and artificial intelligence (ChatGPT-4).
Methods: An observational, descriptive, cross-sectional study with a control group was conducted on a prospective cohort of patients treated for tibial plateau fractures between 2020 and 2024. Anteroposterior radiographs were blindly evaluated by three groups—three knee surgeons, three resident physicians., and ChatGPT-4—with fractures classified according to the Schatzker system. The reference standard was computed tomography (CT). The interobserver agreement was assessed using the Kappa statistic for fracture detection and the Ciccetti weighted Kappa for fracture classification, with a 95% confidence interval. A significance level of $p < 0.01$ was established.
Results: A total of 387 radiographs were included, of which 129 showed tibial plateau fractures (classified according to Schatzker as follows: 7 type I, 28 type II, 5 type III, 16 type IV, 21 type V, and 52 type VI) and 258 were without fracture. The AI demonstrated the highest accuracy in fracture detection, achieving an absolute agreement of 99.5% and a Kappa of 0.98 (95% CI: 0.97–1.00, $p < 0.001$), compared to 97% ($K = 0.93$, 95% CI: 0.91–0.95, $p < 0.001$) for knee surgeons and 93% ($K = 0.848$, 95% CI: 0.81–0.88, $p < 0.001$) for residents. In terms of interobserver variability for fracture diagnosis, the AI showed greater consistency than the human evaluators; however, for fracture classification, knee surgeons achieved a higher weighted Kappa (0.616, 95% CI: 0.554–0.679, $p < 0.001$) compared to the AI (0.612, 95% CI: 0.502–0.722, $p < 0.001$) and residents (0.572, 95% CI: 0.510–0.635, $p < 0.001$).
Conclusions: Artificial intelligence demonstrated notable accuracy in the detection of tibial plateau fractures, outperforming both residents and attending physicians in this specific task. However, in the classification of fractures using the Schatzker system, attending physicians achieved higher accuracy. These findings suggest that AI may serve as a valuable support tool in the diagnostic process, particularly in its early stages, complementing—but not replacing—the clinical judgment and experience of healthcare professionals.
Level of evidence: level III. Diagnostic. Cross-sectional descriptive study with control group.

Introducción

La meseta tibial es una de las regiones que soporta mayor carga durante las actividades diarias. Las fracturas de meseta tibial representan aproximadamente el 1% de todas las fracturas en adultos y hasta el 8% en individuos mayores de 65 años[1]. En años recientes se ha observado un incremento en su incidencia, lo que se asocia a una mayor morbilidad y a un impacto económico significativo en los sistemas de salud[1,2]. Un diagnóstico preciso resulta esencial para definir el tratamiento adecuado[3,4]; sin embargo, la tasa de error diagnóstico sigue siendo elevada, en gran parte debido a interpretaciones erróneas de las radiografías[5]. Para orientar el tratamiento, se han desarrollado distintas clasificaciones, siendo la de Schatzker una de las más utilizadas. Esta se basa en el análisis de radiografías simples, lo que puede limitar su precisión diagnóstica en fracturas complejas[6]. Diversos estudios han señalado que entre el 2% y el 9% de las fracturas tibiales no se detectan en las radiografías realizadas en urgencias, lo que conduce a diagnósticos equivocados y a un aumento en la necesidad de reevaluaciones clínicas o pruebas complementarias. Aunque las radiografías constituyen la herramienta estándar para la evaluación inicial, el análisis de las imágenes está sujeto a errores humanos, atribuibles a factores como la fatiga profesional, la elevada carga de trabajo, la sutileza de ciertas fracturas y la limitada experiencia en algunos casos[7].

En este contexto, la inteligencia artificial (IA) ha experimentado avances significativos en diversas disciplinas, incluida la Cirugía Ortopédica y Traumatología. Se emplea para mejorar la precisión diagnóstica, la formación de cirujanos mediante simulaciones virtuales de procedimientos quirúrgicos y el desarrollo de algoritmos predictivos que permiten anticipar complicaciones y resultados. Además, la IA está transformando la práctica clínica gracias a su capacidad para identificar patrones en grandes volúmenes de datos, lo que contribuye a reducir la variabilidad en los diagnósticos. Asimismo, podría desempeñar un papel clave en la optimización de la rehabilitación posquirúrgica, facilitando el seguimiento y la mejora de los tratamientos[8–13].

En los últimos años, la IA ha evolucionado hacia modelos generativos capaces no solo de resolver problemas, sino también de aprender de los datos y generar contenido original, como textos, imágenes y videos[14]. Un ejemplo clave de esta evolución es ChatGPT (Generative Pre-trained Transformer), un chatbot desarrollado en 2022 por OpenAI (OpenAI, LLC, San Francisco, California, EE. UU.) y programado en Python. Este sistema ha mejorado significativamente su capacidad para interactuar de manera fluida y natural con los usuarios, respondiendo a preguntas complejas en diversos campos. Además, las versiones más recientes de ChatGPT son multimodales, lo que permite procesar tanto texto como imágenes para generar respuestas, ampliando así sus aplicaciones desde la creación de contenido educativo hasta el desarrollo de modelos predictivos en el ámbito médico[10].

La IA conversacional se fundamenta en 3 componentes clave: el aprendizaje automático, el procesamiento de grandes volúmenes de datos y el procesamiento de lenguaje natural. El aprendizaje automático *(machine learning)* permite que la IA aprenda y se perfeccione a partir de la experiencia y la interacción, sin requerir programación explícita. Este tipo de aprendizaje se basa en algoritmos que analizan extensos conjuntos de datos para identificar patrones y formular predicciones o tomar decisiones, incluso sin intervención humana.

Las redes neuronales, que imitan el funcionamiento del cerebro humano, son fundamentales en este proceso al habilitar el aprendizaje profundo *(deep learning)*, lo que permite procesar datos no estructurados como imágenes, textos o sonidos. Este enfoque resulta esencial para tareas complejas, tales como el reconocimiento de voz e imágenes. Por ejemplo, la versión 3 de ChatGPT fue entrenada con 175.000 millones de parámetros, lo que le permite generar respuestas basadas en modelos predictivos previamente construidos a partir de grandes volúmenes de datos, en lugar de realizar búsquedas en tiempo real. Algunos expertos sugieren que el término 'aprendizaje estadístico computacional' podría ser más adecuado que 'inteligencia artificial' para describir este enfoque, ya que se fundamenta en estadísticas y patrones derivados de vastos conjuntos de datos. No obstante, esta tecnología presenta limitaciones, pues puede producir errores o 'alucinaciones' cuando no procesa adecuadamente los datos de entrada o los patrones estocásticos. Esto subraya la necesidad de una supervisión humana constante y de validar los resultados generados por los sistemas de IA. Además, el fenómeno GIGO *(Garbage In, Garbage Out)* refuerza que la calidad de los resultados depende directamente de la calidad de los datos utilizados en el entrenamiento, lo que enfatiza la importancia de contar con datos precisos y libres de sesgo[15].

El procesamiento de lenguaje natural (PLN) es otro pilar clave en la evolución de la IA conversacional, ya que facilita la interacción entre los sistemas de IA y el lenguaje humano. El PLN permite que las máquinas comprendan, interpreten y generen respuestas que imiten la

comunicación humana, lo que les permite interactuar de manera más fluida y natural con los usuarios. Este componente no solo facilita una mejor comprensión de los comandos de los usuarios, sino que también permite a la IA recordar conversaciones pasadas, lo que mejora la calidad y la continuidad de la interacción. Esto tiene un impacto directo en la mejora de la calidad de la atención médica y, en general, en una mayor eficiencia de los procesos clínicos.

El objetivo de este estudio fue analizar la precisión diagnóstica y la clasificación de fracturas de meseta tibial en 3 grupos: facultativos especialistas de área, médicos internos residentes e IA (ChatGPT). Se pretende determinar cómo la integración de la IA puede mejorar el diagnóstico y la clasificación de fracturas tibiales en este entorno clínico. La hipótesis operativa planteaba que el grupo humano presentaría una mayor precisión en el diagnóstico y la clasificación de las fracturas de meseta tibial en comparación con la IA (ChatGPT)[16,17].

Material y métodos

Se llevó a cabo un estudio observacional descriptivo y transversal, con grupo control, sobre una cohorte retrospectiva de pacientes atendidos en el Hospital Universitario Costa del Sol (HUCS) entre 2020 y 2024 por fractura de meseta tibial. Este proyecto fue aprobado por el Comité de Ética y la Unidad de Investigación del HUCS (número 147-01-2025, de 30 de enero de 2025). Para su correcta elaboración se siguieron las guías STARD *(Standard for Reporting of Diagnostic Accuracy)* y se cumplió la Declaración de Helsinki. El carácter retrospectivo de la serie no requirió la obtención de consentimiento informado.

Los pacientes fueron identificados en la base de datos hospitalaria general, utilizando la Clasificación Internacional de Enfermedades (ICD-10) para el diagnóstico de fractura de tibia de extremo superior proximal (S82.10-S82.15). Las radiografías (Rx) sin fractura fueron obtenidas tras el análisis de pacientes con diagnóstico de artrosis primaria de rodilla (M17.0-12) y por procedimiento de menisectomía (0 SBC), analizando en su historia clínica que carecían de antecedente traumático reciente. La recolección de datos se realizó a través de los programas DIRAYA (software abierto para administraciones públicas, Java EE, Visual Basic, INDRA, Servicio Andaluz de Salud), HP Doctor® (Hewlett Packard, Palo Alto, CA, EE.UU., 2010) y HCIS *(Health Care Information System)*. Aunque no se establecieron criterios de edad, se consideró como criterio de exclusión la presencia de fisis de crecimiento abierta. Otros criterios de exclusión incluyeron casos con procesos patológicos asociados que afectaran la rodilla (tumores e infecciones), antecedentes de cirugía previa con material de osteosíntesis o prótesis, y la ausencia de pruebas complementarias (Rx, TC, etc.). Las imágenes fueron analizadas a través del sistema PACS *(Picture Archiving and Communication System)*, Carestream (Health Spain, S.A., 2016).

Se tuvo en consideración la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, así como el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024, que establece normas armonizadas en materia de IA.

El estándar de referencia (*gold* estándar) en el diagnóstico y la clasificación de las fracturas de meseta fue la tomografía computarizada (TAC). Las radiografías de rodilla sin fractura no contaban con TAC, ya que correspondían a pacientes con dolor de rodilla cuyo seguimiento clínico descartó la presencia de fractura. Estos pacientes, con radiografía normal de rodilla atendidos en el mismo período, se agruparon en una proporción 2:1 en relación con los pacientes con fractura. El grupo humano estuvo compuesto por 3 facultativos especialistas de área con más de 5 años de experiencia y 3 médicos internos residentes en formación. La IA seleccionada fue *Chat Generative Pre-trained Transformer* (ChatGPT-4o).

El investigador principal anonimizó de forma irreversible las imágenes (radiografías anteroposteriores) y las transformó en formato JPEG. Para el grupo humano, las imágenes se subieron a un formulario de Google, donde quedaron almacenadas de forma segura. A través de este cuestionario, cada evaluador accedió individualmente para clasificar las imágenes. En el caso de la IA, las imágenes se introdujeron una a una en la propia interfaz de ChatGPT. Se formularon dos preguntas consecutivas:

1. Identifica si existe fractura en la meseta tibial.
2. Si hay fractura, clasifícala según la clasificación de Schatzker (I, II, III, IV, V o VI).

Se utilizó el siguiente *prompt* (instrucción o pregunta formulada al sistema de IA para generar una respuesta) para facilitar la respuesta de la IA:

> Eres un traumatólogo especializado en cirugía ortopédica y traumatología. Te proporcionaré una imagen de radiografía en proyección anteroposterior (AP) de una rodilla. Tu tarea es analizarla e indicar si existe una fractura de meseta tibial. En caso afirmativo, clasifícala

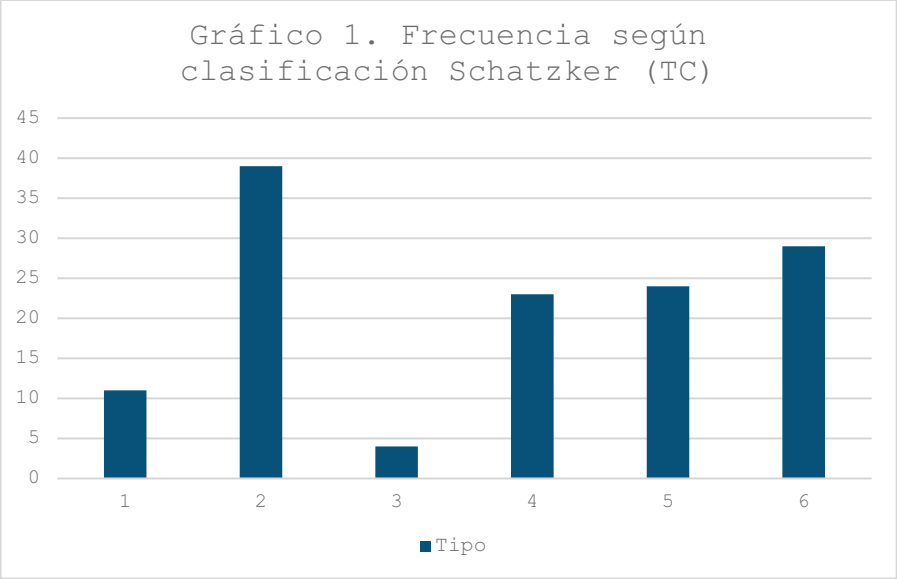

Figura 1. Frecuencia según grado/tipo de clasificación de Schatzker, valorada en tomografía axial computarizada.

Tabla 1
Comparativa del grado de acuerdo absoluto y el valor Kappa respecto al diagnóstico de fractura

	IA		Acuerdo absoluto (%)	Kappa	IC95%		p
Fractura	Ausencia	Presencia			Inferior	Superior	
Ausencia	256	2	99,5	0,988	0,972	1,000	< 0,001
Presencia	0	129					
	Residentes		Acuerdo absoluto (%)	Kappa	IC95%		p
Fractura	Ausencia	Presencia			Inferior	Superior	
Ausencia	710	64	93,0	0,848	0,816	0,880	< 0,001
Presencia	17	370					
	Adjuntos		Acuerdo absoluto (%)	Kappa	IC95%		p
Fractura	Ausencia	Presencia			Inferior	Superior	
Ausencia	745	17	97,0	0,932	0,910	0,955	< 0,001
Presencia	17	356					

IA: inteligencia artificial; IC: intervalo de confianza.

según el sistema de Schatzker. Responde únicamente con el número correspondiente al tipo de fractura (I-VI), o con 'sin fractura' si no se identifica ninguna.

Para reforzar la comprensión de la tarea, se incluyó una breve descripción de los tipos de la clasificación de Schatzker:

- Tipo I: Fractura sin depresión del platillo tibial.
- Tipo II: Fractura del platillo lateral con depresión.
- Tipo III: Fractura por hundimiento (puro).
- Tipo IV: Fractura del platillo medial.
- Tipo V: Fractura que afecta ambas mesetas tibiales.
- Tipo VI: Fractura de la meseta tibial con extensión a la diáfisis.

El análisis estadístico se realizó mediante los programas SPSS® (IBM® V28) y EpiDat 3.1. Se evaluó el rendimiento diagnóstico de la IA, de los médicos internos residentes y de los facultativos, comparándolo con el estándar de referencia (TAC). Para ello, se aplicó el índice Kappa para la detección de fracturas y el Kappa ponderado por Ciccetti para su clasificación, con intervalos de confianza al 95%. Además, se calculó el grado de acuerdo absoluto y se estableció un nivel de significación estadística de $p < 0,01$, dada la realización de múltiples comparaciones.

Resultados

Se evaluaron un total de 387 radiografías, distribuidas en dos grupos: 129 correspondientes a pacientes con fracturas de meseta tibial y 258 a pacientes sin fractura. Según la clasificación de Schatzker establecida mediante TC, se encontraron: 7 fracturas tipo I, 28 tipo II, 5 tipo III, 16 tipo IV, 21 tipo V y 52 tipo VI (fig. 1).

Respecto a la detección de fracturas (tabla 1), la IA presentó la mayor precisión diagnóstica con un acuerdo absoluto del 99,5% y un Kappa de 0,98 (IC 95%, 0,97-1,00, $p < 0,001$), frente a un 97% de los facultativos ($K = 0,93$, IC95% 0,91-0,95, $p < 0,001$) y un 93% de los residentes ($K = 0,848$, IC 95% 0,81-0,88, $p < 0,001$).

En cuanto a la clasificación de las fracturas según la escala de Schatzker (tabla 2), el acuerdo absoluto fue del 84,8% para los facultativos, 84,5% para la IA y 82,8% para los residentes. El índice Kappa fue de 0,616 para los facultativos (IC95%, 0,554-0,679, $p < 0,001$), 0,612 para la IA (IC95%, 0,502-0,722, $p < 0,001$), 0,572 (IC95%, 0,510-0,635, $p < 0,001$) para los residentes, lo que sugiere que los adjuntos presentaron un mejor desempeño en la clasificación de fracturas. Todos los valores fueron estadísticamente significativos ($p < 0,001$).

Las matrices de confusión revelaron que la IA alcanzó una mayor precisión en la clasificación de fracturas de grado VI, con 42 casos correctamente clasificados, mientras que los residentes presentaron mayor dificultad en los grados intermedios. Asimismo, la IA tuvo problemas para determinar correctamente los tipos I y IV. Los especialistas mostraron un desempeño superior a los residentes, con una menor dispersión de errores en la clasificación (tabla 3). En la tabla 3, se expresan los resultados de la IA multiplicados por 3 para que puedan ser comparados con el sumatorio total de adjuntos y residentes.

Tabla 2
Comparativa del grado de acuerdo absoluto y el valor Kappa respecto a clasificación de la fractura

Evaluador	Acuerdo absoluto (%)	Kappa	IC95% inferior	IC95% superior	p
IA	84,5	0,612	0,502	0,722	< 0,001
Residentes	82,8	0,572	0,510	0,635	< 0,001
Adjuntos	84,8	0,616	0,554	0,679	< 0,001

IA: inteligencia artificial; IC: intervalo de confianza.

En resumen, la IA demostró un rendimiento superior en la detección de fracturas (presencia/ausencia), demostrando una alta capacidad para diferenciar entre fractura y no fractura. No obstante, respecto a la clasificación de los tipos de fracturas, los facultativos fueron más precisos y consistentes que la IA, lo que subraya la importancia de la experiencia clínica en la evaluación detallada según la escala de Schatzker. Además, se observó que este chatbot no está exento de limitaciones técnicas. En concreto, a partir de la décima imagen aproximadamente, el sistema presentó cierta dificultad para procesar múltiples radiografías de forma consecutiva, lo que obligó a realizar el análisis de forma individualizada imagen por imagen. Esta circunstancia supuso una ralentización del proceso, aunque sin afectar al rendimiento diagnóstico ni a los resultados estadísticos obtenidos (figs. 2 y 3).

La base de datos anonimizada utilizada en este estudio está disponible públicamente en un repositorio de acceso abierto en Mendeley Data en el DOI: 10.17632/46ff95x62f.1

Discusión

La fractura de meseta tibial es una lesión compleja frecuentemente causada por mecanismos de alta energía. Los pacientes de edad avanzada pueden presentar este tipo de lesiones ante traumatismos de baja energía, como caídas de propia altura[1,2]. El diagnóstico por radiología simple no siempre es suficiente, pues pequeñas fracturas, no desplazadas, ocultas o con escaso hundimiento, pueden llegar a oca-

Diagnóstico: No hay fractura.

Diagnóstico: **Fractura Schatzker tipo IV.**

Figura 2. Observamos un error cometido por la IA, tras varias imágenes analizadas. Se debe tener en cuenta que la IA puede tener episodios de colapso o mal funcionamiento.

Tabla 3
Registro de errores según grado de la clasificación de Schatzker

Schatzker Grado	IA Grado					
	1	2	3	4	5	6
1	6	9	3	0	0	3
2	6	63	3	3	3	6
3	0	6	9	0	0	0
4	3	9	0	30	6	0
5	3	24	3	3	24	6
6	6	6	6	0	12	126

Schatzker Grado	Residentes Grado					
	1	2	3	4	5	6
1	11	6	0	0	1	2
2	14	45	14	0	2	2
3	3	3	6	0	0	0
4	1	2	0	36	4	1
5	7	11	7	14	14	8
6	7	11	1	13	30	94

Schatzker Grado	Adjuntos Grado					
	1	2	3	4	5	6
1	11	4	0	0	0	1
2	9	53	10	0	2	1
3	0	6	5	1	0	0
4	0	1	0	36	2	2
5	7	13	6	7	15	12
6	2	16	0	9	28	97

La tabla IA Grado corregida × 3. Las tablas de facultativos y residentes recopilan el sumatorio de los errores de los 3 integrantes de cada grupo. Es importante indicar la capacidad de acierto de la IA en el tipo VI, a diferencia de los tipos I y IV.
IA: inteligencia artificial.

sionar un diagnóstico erróneo[8]. Además, en nuestra práctica clínica diaria, estos pacientes son inicialmente atendidos por profesionales no especialistas en traumatología, en salas saturadas de Urgencias, o por traumatólogos en formación, lo cual dificulta aún más el correcto diagnóstico y, por ende, el pronóstico de la lesión[7]. En este sentido, la IA puede llegar a ser valorada como una herramienta de apoyo en el diagnóstico. Nuestro estudio apoya a la bibliografía emergente[8,10–12] a favor del uso de IA en el diagnóstico de fracturas de meseta tibial (IA K = 0,98 > Especialistas en cirugía ortopédica y traumatología [COT] K = 0,93).

La clasificación de las fracturas del platillo tibial ha sido fundamental para la decisión del tratamiento quirúrgico y para establecer un pronóstico. Uno de los sistemas de clasificación más utilizados fue el propuesto por Schatzker en 1974[18]. A pesar de su popularidad, la clasificación ha sido criticada por su limitada reproducibilidad, su variabilidad interobservador y por la descripción en un solo plano (AP) de radiografía simple[19]. El uso de radiografías simples, especialmente en proyecciones AP, puede subestimar la complejidad de las fracturas de la meseta tibial. La identificación de fragmentos óseos pequeños, el hundimiento articular y la extensión real de las fracturas en el plano coronal son errores comúnmente cometidos, lo que lleva a una clasificación incorrecta. Asimismo, la naturaleza biplanar de la proyección AP radiográfica limita la capacidad para evaluar adecuadamente la extensión tridimensional de la fractura, lo que es especialmente problemático en fracturas tipo II-III y V-VI. La literatura recoge que la variabilidad intra- e interobservador utilizando radiografías simples fue moderada (κ = 0,40-0,60). Esto sugiere que la dependencia exclusiva de radiografías AP no es suficiente para tomar decisiones clínicas óptimas[6]. Estudios recientes han demostrado que las modalidades avanzadas de imagen, como la tomografía computarizada (TC) y la reconstrucción en 3 D, pueden mejorar considerablemente la precisión diagnóstica y la planificación quirúrgica[19].

En base a lo descrito, en nuestro estudio se ha usado la TC como *gold* estándar en el diagnóstico y clasificación de las fracturas, mientras que los especialistas, el personal en formación y la IA (ChatGPT) sólo han valorado la proyección AP de la radiografía simple. El objetivo fue ver la capacidad de la IA en el diagnóstico y clasificación de este tipo de fracturas en radiografía simple y contrastar la variabilidad respecto al humano. Se observó un resultado superior de la IA (K = 0,98) en el diagnóstico de fractura respecto al especialista (K = 0,9) y al personal en formación (K = 0,85). En contraposición, la capacidad de clasificación de la IA fue ligeramente inferior. A pesar de estos resultados, la precisión en la clasificación de este tipo de fracturas por especialistas e IA (K = 0,60) se asemeja a lo previamente publicado, lo cual indica las carencias que presenta este tipo de clasificación.

Bousson et al.[20] describieron una precisión del 90,1% en el diagnóstico de fracturas de todo el cuerpo, sometiendo a 3 sistemas de IA al análisis de 1500 radiografías. Concluyeron que la parte del cuerpo es importante a la hora de ser analizada por la IA, siendo más precisa en las regiones de rodilla y pie.

Liu et al.[8] desarrollaron una IA propia, la cual entrenaron con radiografías previamente. Estos autores alcanzaron una precisión del 91%, similar a los expertos que participaron. Observaron que el diagnóstico de las fracturas fue 16 veces más rápido respecto a los especialistas, los cuales precisaron de más información para llegar a un diagnóstico.

Figura 3. Capacidad de diagnóstico de múltiples imágenes. Se observan errores a partir de más de 10 imágenes.

Su IA entrenada fue superior a una IA genérica como CHAT GPT 4 o, si bien estos autores no incluyeron radiografías sin fractura. También describen como limitación que no han incluido proyecciones laterales y que hubiera sido importante haber analizado una clasificación de fractura. En nuestro estudio, con una IA sin entrenamiento previo, pero codificando correctamente los *prompts*, observamos que la IA alcanza hasta un 98% de precisión en el diagnóstico de fractura. Además, han sido incluidas imágenes con y sin fractura. Las proyecciones laterales no fueron incluidas pues uno de los objetivos fue realizar un análisis crítico de la clasificación de Schatzker.

Mohammadi et al.[21] observaron que el radiólogo tiene una mayor sensibilidad en el diagnóstico, frente a una mayor especificidad por parte de ChatGPT. El Likelihood ratio + y – era superior en el caso del ChatGPT. Concluyen que la capacidad diagnóstica de la IA se asemeja a la del médico. Por ello, la IA puede ser considerada una herramienta que aporte información extra a la hora de valorar una radiografía.

Actualmente, existen diversas compañías tecnológicas que han desarrollado nuevos sistemas chatbots de IA genérica. Entre estos destacan Bing (Microsoft Corporation, Redmond, Washington, EE. UU.), Bard de Google (Google LLC, Mountain View, California, EE. UU.), Perplexity (Perplexity AI, 2022, Aravind Srinivas) o DeepSeek (DeepSeek LLM, High-Flyer, Hangzhou, China). Este estudio ha utilizado ChatGPT 4 o (OpenAI, LLC, San Francisco, California, EE. UU.) al ser el sistema de IA más extendido en la actualidad. Aún existe escasa evidencia que contraste la capacidad diagnóstica y de clasificación de los diferentes sistemas de IA, aunque algunos autores no encuentran grandes diferencias[20].

La incorporación de tecnologías como ChatGPT en la práctica médica, aunque prometedora, plantea una serie de retos éticos y legales. A pesar de las ventajas que ofrece en términos de precisión diagnóstica y eficiencia, surgen interrogantes sobre la responsabilidad en caso de errores[13]. Si un sistema de IA comete un diagnóstico erróneo, ¿quién debe asumir la culpa: el desarrollador, el médico que lo emplea o el propio sistema? Este tipo de dudas genera incertidumbre en cuanto a la toma de decisiones clínicas y la confianza de los pacientes en los profesionales de salud. Además, el uso de la IA podría fomentar una dependencia de la tecnología que reduzca la autonomía del médico, lo cual puede poner en riesgo el juicio clínico. Si bien la IA puede reducir la variabilidad diagnóstica, su función debe ser complementaria al conocimiento y la experiencia del profesional de la salud, no un reemplazo de estos. También es crucial considerar que los algoritmos podrían no identificar todos los problemas asociados a las fracturas, como los tumores óseos, lo que refuerza la necesidad de un enfoque equilibrado[11,13,20,22,23]. En definitiva, estos aspectos deben ser cuidadosamente evaluados antes de su implementación en el entorno clínico. En nuestro estudio, se observa que la IA tiene limitaciones en cuanto a rendimiento, teniendo serias dificultades a la hora de procesar más de 10 imágenes a la vez y con períodos de inactividad obligada por parte del sistema. Por tanto, este gran avance aún requiere de mejoras para optimizar su prestación, lo cual no evita que sea una herramienta cuya precisión avanza con rapidez.

Nuevas líneas de investigación son necesarias para analizar la capacidad de la IA en la reconstrucción volumétrica a partir de proyecciones simples, en la decisión terapéutica (tipos de implantes, abordajes…) y en la valoración pronóstica, que permita planificar con exactitud cada intervención.

Sin embargo, este estudio no está exento de limitaciones. ChatGPT es un modelo de lenguaje especializado en el diálogo, cuyo procesamiento es ajustado mediante técnicas de aprendizaje supervisadas y de refuerzo. No obstante, este aprendizaje es ampliado por plugins que permiten a la IA el acceso a internet. Aunque tiene constantes actualizaciones, las fracturas de meseta tibial es un tema especializado y de un ámbito profesional muy reducido, cuya información presente es aún escasa. Esto da pie a que especialistas en este tipo de patología puedan llegar a presentar una superioridad en el diagnóstico y la clasificación. El tamaño muestral podría ser considerado como una limitación, no obstante, fueron reclutados todos los casos de fractura de meseta tibial atendidos en un único centro durante los últimos 5 años, estableciendo una relación 1:2 con pacientes con radiografía de rodilla sin traumatismo y sin antecedentes de fractura. Otra limitación importante es la ausencia de análisis de la radiografía lateral, si bien fue intencionado para plantear un desafío diagnóstico tanto para la IA como para los evaluadores humanos, al limitar la información disponible a una sola proyección. Además, varios estudios publicados[8,21] utilizan únicamente la proyección AP para el diagnóstico y clasificación de fracturas de meseta tibial.

Conclusión

La IA (ChatGPT-4o) puede constituir una herramienta complementaria en el diagnóstico inicial y en la clasificación de fracturas de meseta tibial. No obstante, su uso no sustituye la valoración clínica integral ni el juicio del especialista, incluyendo la elaboración de una historia clínica completa. Los chatbots deben entenderse como sistemas de apoyo, capaces de aportar información adicional que contribuya a optimizar la práctica clínica. Por otro lado, la clasificación de Schatzker presenta una concordancia interobservador baja a moderada, por lo que sus resultados deben interpretarse con cautela, especialmente cuando se basan exclusivamente en radiografías simples.

Nivel de evidencia

Nivel de evidencia III.

Consideraciones éticas

- Aprobación por el Comité de Ética de Investigación Costa del Sol, CEIC acreditado y constituido conforme a los requisitos recogidos en el Decreto 8/2020 (202599900951259).
- Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, así como el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024, que establece normas armonizadas en materia de inteligencia artificial.

Financiación

No se ha precisado financiación para este estudio.

Conflicto de intereses

Los autores declaran que no han recibido apoyo financiero ni material en la investigación, autoría ni publicación de este artículo.

Declaro que no tengo conflicto de interés en este estudio.

Bibliografía

1. Herteleer M, Van Brandt C, Vandoren C, Nijs S, Hoekstra H. Tibial plateau fractures in Belgium: epidemiology, financial burden and costs curbing strategies. *Eur J Trauma Emerg Surg*. 2022;48:3643–3650.
2. Bormann M, Neidlein C, Gassner C, et al. Changing patterns in the epidemiology of tibial plateau fractures: a 10-year review at a level-I trauma center. *Eur J Trauma Emerg Surg*. 2023;49:401–409.
3. Samsami S, Pätzold R, Winkler M, Herrmann S, Augat P. The effect of coronal splits on the structural stability of bi-condylar tibial plateau fractures: a biomechanical investigation. *Arch Orthop Trauma Surg*. 2020;140:1719–1730.
4. Schatzker J, Kfuri M. Revisiting the management of tibial plateau fractures. *Injury*. 2022;53:2207–2218.
5. Kiel CM, Mikkelsen KL, Krogsgaard MR. Why tibial plateau fractures are overlooked. *BMC Musculoskelet Disord*. 2018;19:244.
6. Millar SC, Arnold JB, Thewlis D, Fraysse F, Solomon LB. A systematic literature review of tibial plateau fractures: What classifications are used and how reliable and useful are they? *Injury*. 2018;49:473–490.
7. Pinto A, Reginelli A, Pinto F, et al. Errors in imaging patients in the emergency setting. *Br J Radiol*. 2016;89:20150914.
8. Liu P, Zhang J, Xue M, et al. Artificial Intelligence to Diagnose Tibial Plateau Fractures: An Intelligent Assistant for Orthopedic Physicians. *Curr Med Sci*. 2021;41:1158–1164.
9. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2024;66:73–79.
10. Tustumi F, Andreollo NA, Aguilar-Nascimento JEde. Future of the language models in healthcare: the role of ChatGPT. *Arq Bras Cir Dig*. 2023;36.
11. Bhatnagar A, Kekatpure AL, Velagala VR, Kekatpure A. A Review on the Use of Artificial Intelligence in Fracture Detection. *Cureus*. 2024.
12. Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS. Artificial intelligence in orthopaedic surgery. *Bone Joint Res*. 2023;12:447–454.
13. Oosterhoff JHF, Doornberg JN. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev*. 2020;5:593–603.
14. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine*. 2019;46:27–29.
15. Murphy MP, Brown NM, Synthesis: CORR. When Should the Orthopaedic Surgeon Use Artificial Intelligence Machine Learning, and Deep Learning? *Clin Orthop Relat Res*. 2021;479:1497–1505.
16. Canillas del Rey F, Canillas Arias M. Explorando el potencial de la inteligencia artificial en traumatología: respuestas conversacionales a preguntas específicas. *Rev Esp Cir Ortop Traumatol*. 2025;69:38–46.
17. Kuo RYL, Harrison C, Curran TA, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology*. 2022;304:50–62.
18. Schatzker J. Compression in the surgical treatment of fractures of the tibia. *Clin Orthop Relat Res*. 1974:220–239.
19. Castiglia M, Nogueira-Barbosa M, Messias A, et al. The Impact of Computed Tomography on Decision Making in Tibial Plateau Fractures. *J Knee Surg*. 2018;31:1007–1014.
20. Bousson V, Attané G, Benoist N, et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. *Acad Radiol*. 2023;30:2118–2139.
21. Mohammadi M, Parviz S, Parvaz P, Pirmoradi MM, Afzalimoghaddam M, Mirfazaelian H. Diagnostic performance of ChatGPT in tibial plateau fracture in knee X-ray. *Emerg Radiol*. 2024;32:59–64.
22. Millán-Billi A, Gómez-Masdeu M, Ramírez-Bermejo E, Ibañez M, Gelber PE. What is the most reproducible classification system to assess tibial plateau fractures? *Int Orthop*. 2017;41:1251–1256.
23. Aedo-Martín D. [Translated article] Artificial intelligence: Future and challenges in modern medicine. *Rev Esp Cir Ortop Traumatol*. 2024;68:T428–T429.