Revista Española de Cirugía Ortopédica y Traumatología
SOCIEDAD ESPAÑOLA DE CIRUGÍA ORTOPÉDICA Y TRAUMATOLOGÍA (SECOT)
1
2021

[Translated article] Reliability of artificial intelligence (ChatGPT) in the diagnosis and classification of tibial plateau fractures

C. Castillejo M. Zapatero JM. Bogallo F. Lorente C. Ortiz J. Romero F. Rivas-Ruiz ML. Bertrand

PII: S1888-4415(25)00199-7

DOI: https://doi.org/doi:10.1016/j.recot.2025.11.017

Reference: RECOT 1646

To appear in: *Revista Espanola de Cirugia Ortopedica y Traumatologia*

Received Date: 12 March 2025

Accepted Date: 11 August 2025

es
fla
RECOT 1615
S1888-4415(25)00154-7
10.1016/j.recot.2025.08.004
unknown 2025
Received: 12-3-2025; **Accepted:** 11-8-2025

Original

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

[Translated article] Reliability of artificial intelligence (ChatGPT) in the diagnosis and classification of tibial plateau fractures

C. Castillejo **0009-0005-6631-8200**[a*]coralcastillejoiniesta@gmail.com, M. Zapatero[a], J.M. Bogallo[a], F. Lorente[a], C. Ortiz[a], J. Romero[a], F. Rivas-Ruiz[b], M.L. Bertrand[a]

[a] <org>Servicio de Cirugía Ortopédica y Traumatología, Hospital Universitario Costa del Sol, Universidad de Málaga, Marbella</org>, <city>Málaga</city>, <country>Spain</country>

a Servicio de Cirugía Ortopédica y Traumatología. Hospital Universitario Costa del Sol, Universidad de Málaga. Marbella, Málaga, Spain

[b] <org>Unidad de Apoyo a la Investigación, Hospital Universitario Costa del Sol, Universidad de Málaga, Marbella</org>, <city>Málaga</city>, <country>Spain</country>

b Unidad de Apoyo a la Investigación. Hospital Universitario Costa del Sol, Universidad de Málaga. Marbella, Málaga, Spain

[*] Corresponding author.

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.

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.

Palabras clave: Fractura de meseta tibial; Inteligencia artificial; ChatGPT

Keywords: Tibial plateau fracture; Artificial Intelligence; ChatGPT

Introduction

The tibial plateau is one of the regions that bears the greatest load during daily activities. Tibial plateau fractures represent approximately 1% of all fractures in adults and up to 8% in individuals over 65 years of age.[1] In recent years, an increase in their incidence has been observed, which is associated with greater morbidity and a significant economic impact on healthcare systems.[1,2] Accurate diagnosis is essential to define appropriate treatment.[3,4] However, the diagnostic error rate remains high, largely due to misinterpretations of radiographs.[5] To guide treatment, various classifications have been developed, with the Schatzker classification being one of the most widely used. This classification is based on the analysis of plain radiographs, which can limit its diagnostic accuracy in complex fractures.[6] Several studies have indicated that between 2% and 9% of tibial fractures are not detected on X-rays taken in the emergency department, leading to misdiagnoses and an increased need for clinical re-evaluations or further testing. Although X-rays are the standard tool for initial assessment, image analysis is subject to human error, attributable to factors such as professional fatigue, heavy workload, the subtle nature of certain fractures, and limited experience in some cases.[7]

In this context, artificial intelligence (AI) has made significant progress in various disciplines, including orthopaedic surgery and traumatology. It is used to improve diagnostic accuracy, train surgeons through virtual simulations of surgical procedures, and develop predictive algorithms that facilitate anticipating complications and outcomes. Furthermore, AI is transforming clinical practice thanks to its ability to identify patterns in large volumes of data, which helps reduce variability in diagnoses. It could also play a key role in optimising post-surgical rehabilitation, facilitating the monitoring and improvement of treatments.[8–13]

In recent years, AI has evolved toward generative models capable not only of solving problems but also of learning from data and generating original content, such as text, images, and videos.[14] A key example of this evolution is ChatGPT (Generative Pre-trained Transformer), a chatbot developed in 2022 by OpenAI (OpenAI, LLC, San Francisco, California, USA) and programmed in Python. This system has significantly improved its ability to interact fluently and naturally with users, answering complex questions in various fields. Furthermore, the latest versions of ChatGPT are multimodal, allowing it to process both text and images to generate responses, thus expanding its applications from the creation of educational content to the development of predictive models in the medical field.[10]

Conversational AI is based on three key components: machine learning, big data processing, and natural language processing. Machine learning allows AI to learn and improve from experience and interaction, without requiring explicit programming. This type of learning relies on algorithms that analyse vast datasets to identify patterns and formulate predictions or make decisions, even without human intervention.

Neural networks, which mimic the workings of the human brain, are fundamental to this process by enabling deep learning, which allows for the processing of unstructured data such as images, text, or sounds. This approach is essential for complex tasks, such as speech and image recognition. For example, version 3 of ChatGPT was trained with 175 billion parameters, allowing it to generate responses based on predictive models previously built from large volumes of data, rather than performing real-time searches. Some experts suggest that the term “computational statistical learning” might be more appropriate than “artificial intelligence” to describe this approach, as it is based on statistics and patterns derived from vast datasets. However, this technology has limitations, as it can produce errors or “hallucinations” when it does not properly process input data or stochastic patterns. This underscores the need for constant human oversight and validation of the results generated by AI systems. Furthermore, the GIGO (Garbage In, Garbage Out) phenomenon reinforces that the quality of the results depends directly on the quality of the data used in training, emphasising the importance of having accurate and unbiased data.[15]

Natural language processing (NLP) is another key pillar in the evolution of conversational AI, as it facilitates interaction between AI systems and human language. Natural Language Processing (NLP) allows machines to understand, interpret, and generate responses that mimic human communication, enabling them to interact more fluidly and naturally with users. This component not only facilitates a better understanding of user commands but also allows AI to remember past conversations, improving the quality and continuity of the interaction. This has a direct impact on improving the quality of healthcare and, more generally, on increasing the efficiency of clinical processes.

The aim of this study was to analyse the diagnostic accuracy and classification of tibial plateau fractures in three groups: specialist physicians, resident physicians, and an AI (ChatGPT). The goal was to determine how the integration of AI can improve the diagnosis and classification of tibial plateau fractures in this clinical setting. The operational hypothesis was that the human group would demonstrate greater accuracy in the diagnosis and classification of tibial plateau fractures compared to the AI (ChatGPT).[16,17]

Material and methods

A descriptive, cross-sectional, observational study with a control group was conducted on a retrospective cohort of patients treated at the Costa del Sol University Hospital (HUCS) between 2020 and 2024 for tibial plateau fractures. This project was approved by the Ethics Committee and the Research Unit of HUCS (number 147-01-2025, dated January 30, 2025). The STARD (Standard for Reporting of Diagnostic Accuracy) guidelines were followed for its proper design, and the Declaration of Helsinki was observed. The retrospective nature of the series did not require obtaining informed consent.

Patients were identified in the general hospital database using the International Classification of Diseases (ICD-10) for the diagnosis of proximal tibial fracture (S82.10-S82.15). Radiographs (X-rays) without fractures were obtained from patients diagnosed with primary knee osteoarthritis (M17.0-12) and those who had undergone meniscectomy (0░SBC), after reviewing their medical records and confirming the absence of recent trauma. Data collection was performed using the DIRAYA programme (open-source software for public administrations, Java EE, Visual Basic, INDRA, Andalusian Health Service), HP Doctor® (Hewlett Packard, Palo Alto, CA, USA, 2010), and HCIS (Health Care Information System). Although no age criteria were established, the presence of an open growth plate was considered an exclusion criterion. Other exclusion criteria included cases with associated pathological processes affecting the knee (tumours and infections), a history of previous surgery with osteosynthesis material or prostheses, and the absence of complementary tests (X-rays, CT scans, etc.). The images were analysed using the PACS (Picture Archiving and Communication System), Carestream (Health Spain, S.A., 2016).

Organic Law 3/2018, of December 5, on the Protection of Personal Data and Guarantee of Digital Rights, as well as Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024, establishing harmonised rules on AI, were taken into consideration.

The gold standard for the diagnosis and classification of plateau fractures was computed tomography (CT). Knee radiographs without fractures did not include CT scans, as they corresponded to patients with knee pain whose clinical follow-up ruled out the presence of a fracture. These patients, with normal knee radiographs seen during the same period, were grouped in a 2:1 ratio compared to patients with fractures. The human team consisted of three specialist physicians with over five years of experience and three resident physicians in training. The selected AI was Chat Generative Pre-trained Transformer (ChatGPT-4o).

The principal investigator irreversibly anonymised the images (anteroposterior radiographs) and converted them to JPEG format. For the human team, the images were uploaded to a Google Form, where they were securely stored. Each evaluator accessed and classified the images individually using this questionnaire. For the AI, the images were entered one by one into the ChatGPT interface. Two consecutive questions were asked:

1. Identify if there is a tibial plateau fracture.

2. If there is a fracture, classify it according to the Schatzker classification (I, II, III, IV, V, or VI).

The following prompt (instruction or question posed to the AI system to generate a response) was used to facilitate the AI's response:

You are a trauma surgeon specialising in orthopaedic surgery and traumatology. I will provide you with an anteroposterior (AP) radiograph of a knee. Your task is to analyse it and indicate whether a tibial plateau fracture is present. If so, classify it according to the Schatzker classification system. Respond only with the number corresponding to the fracture type (I-VI), or with 'no fracture' if none is identified.

To reinforce understanding of the task, a brief description of the Schatzker classification types was included:

- Type I: Fracture of the lateral tibial plateau with no depression.
- Type II: Fracture of the lateral tibial plateau with depression.
- Type III: Pure depression fracture.
- Type IV: Medial tibial plateau fracture
- Type V: Bycondylar fracture.
- Type VI: Fracture with complete dissociation between the tibial plateau (the epiphysis) and the shaft (diaphysis).

Statistical analysis was performed using SPSS® (IBM® V28) and EpiDat 3.1 software. The diagnostic performance of the surgical team, resident physicians, and attending physicians was evaluated and compared to the reference standard (CT scan). The Kappa index was used for fracture detection, and the Ciccetti-weighted Kappa was used for fracture classification, with 95% confidence intervals. Furthermore, the absolute degree of agreement was calculated, and a statistical significance level of $p < .01$ was established, given the multiple comparisons performed.

Results

A total of 387 radiographs were evaluated, distributed into two groups: 129 corresponding to patients with tibial plateau fractures and 258 to patients without fractures. According to the Schatzker classification established by CT scan, the following fractures were found: 7 type I, 28 type II, 5 type III, 16 type IV, 21 type V, and 52 type VI (Fig. 1).

Regarding fracture detection (Table 1), the AI group showed the highest diagnostic accuracy with an absolute agreement of 99.5% and a Kappa of .98 (95% CI, .97–1.00, $p < 0.001$), compared to 97% for physicians (Kappa = .93, 95% CI, .91–.95, $p < .001$) and 93% for residents (Kappa = .848, 95% CI, .81–0.88, $p < .001$).

Regarding fracture classification according to the Schatzker scale (Table 2), absolute agreement was 84.8% for attending physicians, 84.5% for AI, and 82.8% for residents.

The Kappa index was .616 for attending physicians (95% CI, .554–.679, p < .001), .612 for attending physicians (95% CI, .502–.722, p < .001), and .572 (95% CI, .510–0635, p < .001) for residents, suggesting that attending physicians performed better in fracture classification. All values were statistically significant (p < .001).

Confusion matrices revealed that the attending physician achieved greater accuracy in classifying grade VI fractures, with 42 cases correctly classified, while residents had greater difficulty with intermediate grades. The attending physician also struggled to correctly identify types I and IV fractures. Specialists outperformed residents, with a lower dispersion of errors in classification (Table 3). Table 3 presents the attending physician's results multiplied by 3 for comparison with the total scores of attending physicians and residents.

In summary, the AI demonstrated superior performance in fracture detection (presence/absence), showing a high capacity to differentiate between fractures and non-fractures. However, regarding the classification of fracture types, clinicians were more accurate and consistent than the AI, highlighting the importance of clinical experience in detailed assessment according to the Schatzker scale. Furthermore, it was observed that this chatbot is not without technical limitations. Specifically, from approximately the tenth image onward, the system experienced some difficulty processing multiple radiographs consecutively, which necessitated individual image-by-image analysis. This resulted in a slowdown of the process, although without affecting diagnostic performance or the statistical results obtained (Figs. 2 and 3).

The anonymised database used in this study is publicly available in an open access repository on Mendeley Data at DOI: 10.17632/46ff95x62f.1

Discussion

Tibial plateau fractures are complex injuries frequently caused by high-energy mechanisms. Elderly patients may present with this type of injury from low-energy trauma, such as falls from standing height.[1,2] Diagnosis by plain radiography is not always sufficient, as small, non-displaced, occult, or minimally depressed fractures can lead to misdiagnosis.[8] Furthermore, in our daily clinical practice, these patients are initially seen by non-specialist traumatologists in overcrowded emergency departments or by traumatologists in training, which further complicates the correct diagnosis and, consequently, the prognosis of the injury.[7] In this regard, AI can be considered a valuable diagnostic tool. Our study supports the emerging literature[8,10-12] in favour of using the AI classification in the diagnosis of tibial plateau fractures (AI K=.98; Orthopaedic Surgeons [OSS] K=.93).

The classification of tibial plateau fractures has been essential for deciding on surgical treatment and establishing a prognosis. One of the most widely used classification systems was that proposed by Schatzker in 1974.[18] Despite its popularity, the classification has been criticised for its limited reproducibility, interobserver variability, and the single-plane (AP) description of plain radiographs.[19] The use of plain radiographs, especially AP projections, can underestimate the complexity of tibial plateau fractures. The identification of small bone fragments, articular depression, and the true extent of the fractures in the coronal plane are common errors, leading to misclassification. Furthermore, the biplanar nature of the AP radiographic projection limits the ability to adequately assess the three-dimensional extent of the fracture, which is especially problematic in type II-III and V-VI fractures. The literature reports that intra- and interobserver variability using plain radiographs was moderate (κ = .40-.60). This suggests that relying solely on AP radiographs is insufficient for optimal clinical decision-making.[6] Recent studies have demonstrated that advanced imaging modalities, such as computed tomography (CT) and 3D reconstruction, can significantly improve diagnostic accuracy and surgical planning.[19]

Based on the above, our study used CT as the gold standard for diagnosing and classifying fractures, while specialists, residents, and the AI (ChatGPT) only assessed the AP projection of plain radiographs. The objective was to assess AI's ability to diagnose and classify these types of fractures on plain radiographs and to compare its variability with that of humans. AI showed superior results (K=.98) in fracture diagnosis compared to specialists (K=.9) and residents (K=.85). However, the AI's classification ability was slightly lower. Despite these results, the accuracy in classifying these types of fractures by specialists and AI (K=.60) is similar to previously published findings, highlighting the limitations of this type of classification.

Bousson et al.[20] described a 90.1% accuracy rate in diagnosing whole-body fractures, subjecting three AI systems to the analysis of 1,500 radiographs. They concluded that the body part analysed by the AI is important, with the AI being most accurate in the knee and foot regions.

Liu et al.[8] developed their own AI, which they trained using radiographs. These authors achieved a 91% accuracy rate, similar to that of the participating experts. They observed that fracture diagnosis was 16 times faster than that of the specialists, who required more information to reach a diagnosis. Their trained AI outperformed a generic AI such as CHAT GPT 4, although these authors did not include radiographs without fractures. They also noted as a limitation the lack of lateral views and the importance of analysing fracture classification. In our study, using an AI with no prior training but correctly coding the prompts, we observed that the AI achieved up to 98% accuracy in diagnosing fractures. Furthermore, images with and without fractures were included. Lateral projections were not included because one of the objectives was to perform a critical analysis of the Schatzker classification.

Mohammadi et al.[21] observed that radiologists have greater sensitivity in diagnosis, compared to greater specificity for ChatGPT. The positive and negative likelihood ratios were higher for ChatGPT. They concluded that the AI's diagnostic capacity is similar to that of a physician. Therefore, AI can be considered a tool that provides additional information when evaluating radiographs.

Currently, several technology companies have developed new generic AI chatbot systems. Among these are Bing (Microsoft Corporation, Redmond, Washington, USA), Google's Bard (Google LLC, Mountain View, California, USA), Perplexity (Perplexity AI, 2022, Aravind Srinivas), and DeepSeek (DeepSeek LLM, High-Flyer, Hangzhou, China). This study used ChatGPT 4 (OpenAI, LLC, San Francisco, California, USA) as it is currently the most widely used AI system. There is still little evidence comparing the diagnostic and classification capabilities of different AI systems, although some authors find no significant differences.[20]

The incorporation of technologies like ChatGPT into medical practice, while promising, raises a number of ethical and legal challenges. Despite the advantages it offers in terms of diagnostic accuracy and efficiency, questions arise regarding liability in the event of errors.[13] If an AI system makes a misdiagnosis, who should bear the blame: the developer, the physician using it, or the system itself? These kinds of doubts generate uncertainty regarding clinical decision-making and patients' trust in healthcare professionals. Furthermore, the use of AI could foster a dependence on technology that reduces physician autonomy, which can jeopardise clinical judgment. While AI can reduce diagnostic variability, its role should be complementary to the healthcare professional's knowledge and experience, not a replacement for them. It is also crucial to consider that algorithms may not identify all problems associated with fractures, such as bone tumours, reinforcing the need for a balanced approach.[11,13,20,22,23] In short, these aspects must be carefully evaluated before implementation in the clinical setting. Our study shows that AI has performance limitations, experiencing significant difficulties processing more than 10 images simultaneously and requiring periods of system downtime. Therefore, this significant advancement still requires improvements to optimise its performance, although it is a tool whose accuracy is rapidly improving.

Further research is needed to analyse AI's capacity for volumetric reconstruction from simple projections, therapeutic decision-making (types of implants, approaches, etc.), and prognostic assessment, enabling precise planning of each intervention.

However, this study is not without limitations. ChatGPT is a language model specialising in dialogue, whose processing is refined using supervised and reinforcement learning techniques. However, this learning is enhanced by plugins that allow the AI internet access. Although constantly updated, tibial plateau fractures are a specialised topic within a very limited professional field, and current information is still scarce. This allows specialists in this type of pathology to potentially have an advantage in diagnosis and classification. The sample size could be considered a limitation. However, all cases of tibial plateau fracture treated at a single centre over the last 5 years were recruited, establishing a 1:2 ratio of patients with knee radiographs without trauma and without a history of fracture. Another important limitation is the absence of lateral radiograph analysis, although this was intentional to pose a diagnostic challenge for both AI and human evaluators by limiting the available information to a single projection. Furthermore, several published studies[8,21] use only the AP projection for the diagnosis and classification of tibial plateau fractures.

Conclusion

AI (ChatGPT-4o) can be a complementary tool in the initial diagnosis and classification of tibial plateau fractures. However, its use does not replace comprehensive clinical assessment or specialist judgment, including the development of a complete medical history. Chatbots should be understood as support systems capable of providing additional information that contributes to optimising clinical practice. Furthermore, the Schatzker classification exhibits low to moderate interobserver agreement, so its results should be interpreted with caution, especially when based exclusively on plain radiographs.

Level of evidence

Level of evidence III.

Ethical considerations

- Approval by the Costa del Sol Research Ethics Committee, an accredited and constituted CEIC (Research Ethics Committee) in accordance with the requirements of Decree 8/2020 (202599900951259).
- Organic Law 3/2018, of December 5, on the Protection of Personal Data and the guarantee of digital rights, as well as Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024, which establishes harmonised rules in the field of artificial intelligence.

Funding

No funding was required for this study. The authors declare that they have not received any financial or material support for the research, authorship, or publication of this article.

Conflict of interests

I declare that I have no conflict of interest in this study.

References

1. Herteleer M, Van Brandt C, Vandoren C, Nijs S, Hoekstra H. Tibial plateau fractures in Belgium: epidemiology, financial burden and costs curbing strategies. European Journal of Trauma and Emergency Surgery. 2022 Oct 23;48(5):3643–50.
2. Bormann M, Neidlein C, Gassner C, Keppler AM, Bogner-Flatz V, Ehrnthaller C, et al. Changing patterns in the epidemiology of tibial plateau fractures: a 10-year review at a level-I trauma center. European Journal of Trauma and Emergency Surgery. 2023 Feb 3;49(1):401–9.
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 Nov 26;140(11):1719–30.
4. Schatzker J, Kfuri M. Revisiting the management of tibial plateau fractures. Injury. 2022 Jun;53(6):2207–18.
5. Kiel CM, Mikkelsen KL, Krogsgaard MR. Why tibial plateau fractures are overlooked. BMC Musculoskelet Disord. 2018 Dec 21;19(1):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 Mar;49(3):473–90.
7. Pinto A, Reginelli A, Pinto F, Re G Lo, Midiri F, Muzj C, et al. Errors in imaging patients in the emergency setting. Br J Radiol. 2016 May;89(1061):20150914.
8. Liu P ran, Zhang J yao, Xue M di, Duan Y yu, Hu J lang, Liu S xiang, et al. Artificial Intelligence to Diagnose Tibial Plateau Fractures: An Intelligent Assistant for Orthopedic Physicians. Curr Med Sci. 2021 Dec 31;41(6):1158–64.
9. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024 Jan 23;66(1):73–9.
10. Tustumi F, Andreollo NA, Aguilar-Nascimento JE de. FUTURE OF THE LANGUAGE MODELS IN HEALTHCARE: THE ROLE OF CHATGPT. ABCD Arquivos Brasileiros de Cirurgia Digestiva (São Paulo). 2023;36.
11. Bhatnagar A, Kekatpure AL, Velagala VR, Kekatpure A. A Review on the Use of Artificial Intelligence in Fracture Detection. Cureus. 2024 Apr 16;
12. Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS. Artificial intelligence in orthopaedic surgery. Bone Joint Res. 2023 Jul 10;12(7):447–54.
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 Oct;5(10):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 Aug;46:27–9.
15. Murphy MP, Brown NM. CORR Synthesis: When Should the Orthopaedic Surgeon Use Artificial Intelligence, Machine Learning, and Deep Learning? Clin Orthop Relat Res. 2021 Jul;479(7):1497–505.
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 Jan;69(1):38–46.
17. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022 Jul;304(1):50–62.

18. Schatzker J. Compression in the surgical treatment of fractures of the tibia. Clin Orthop Relat Res. 1974;(105):220–39.
19. 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 Mar;49(3):473–90.
20. Castiglia M, Nogueira-Barbosa M, Messias A, Salim R, Fogagnolo F, Schatzker J, et al. The Impact of Computed Tomography on Decision Making in Tibial Plateau Fractures. J Knee Surg. 2018 Nov 14;31(10):1007–14.
21. Bousson V, Attané G, Benoist N, Perronne L, Diallo A, Hadid-Beurrier L, 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 Oct;30(10):2118–39.
22. 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 Nov 30;32(1):59–64.
23. 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 Jun 13;41(6):1251–6.
24. Aedo-Martín D. [Translated article] Artificial intelligence: Future and challenges in modern medicine. Rev Esp Cir Ortop Traumatol. 2024 Jul;68(4):T428–T429.

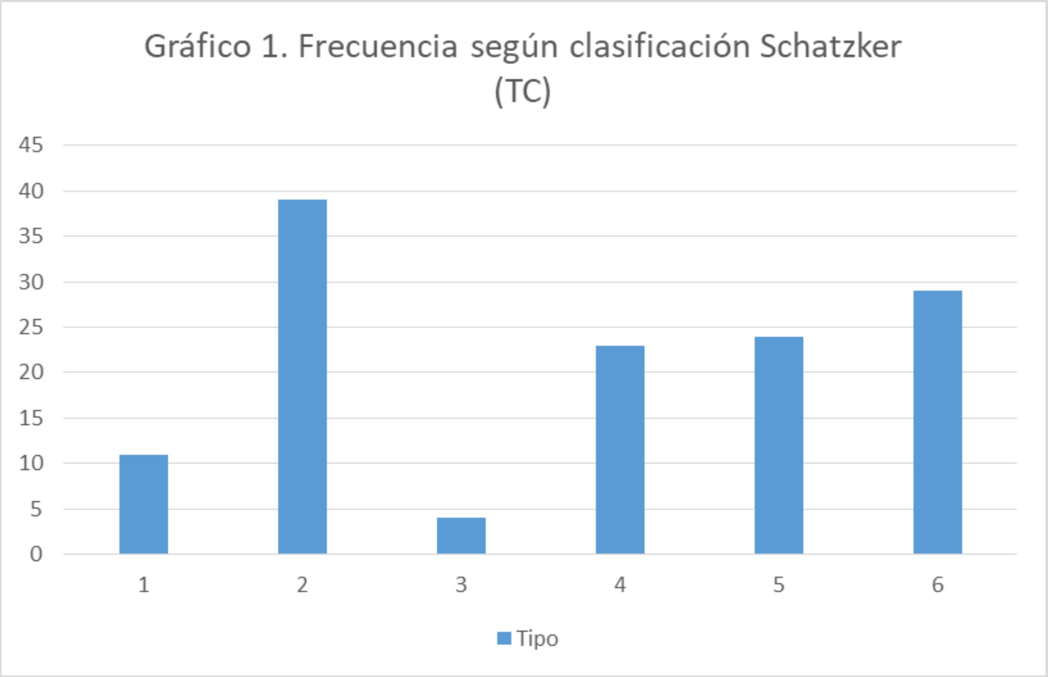
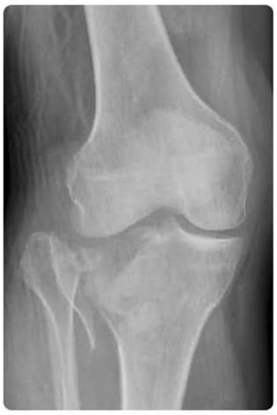
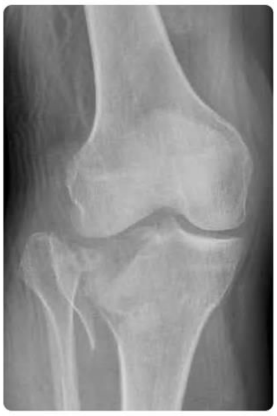


Figure 1. Frequency according to Schatzker classification grade/type, assessed by computed tomography. Gr.1.

Gráfico 1. Frecuencia según clasificación Schatzker (TC)	Graph 1. Frequency according to Schatzker classification (CT)
Tipo	Type



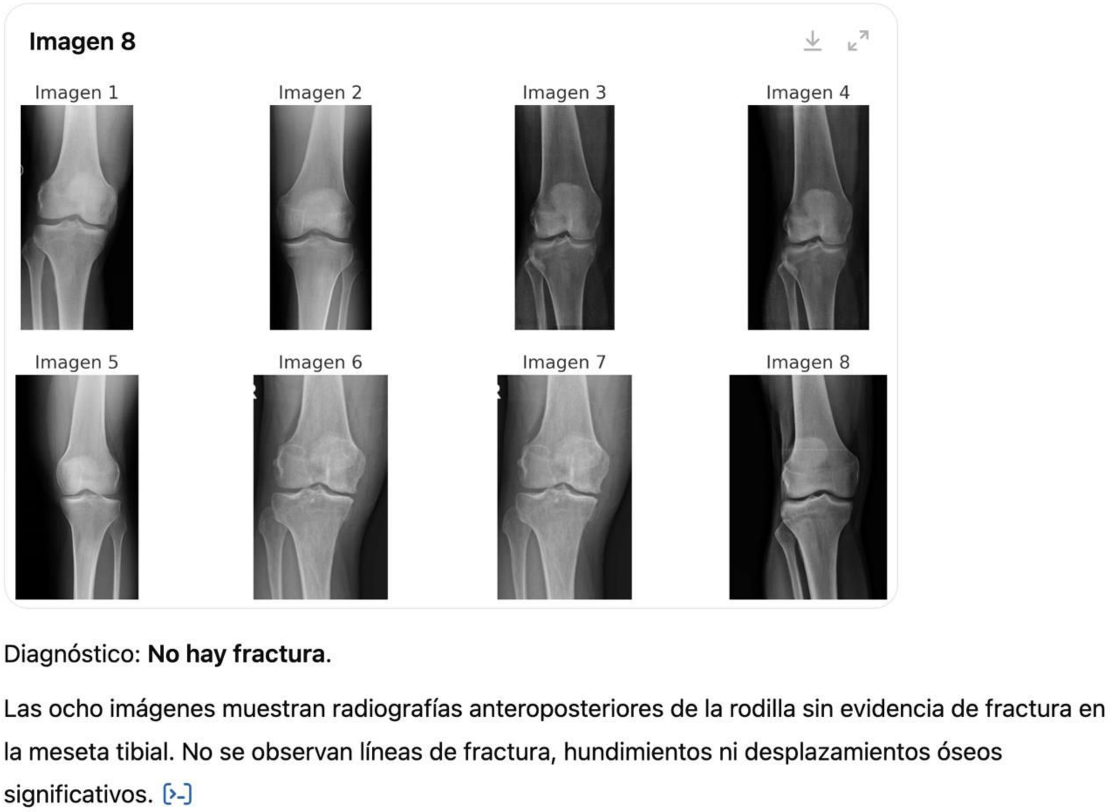
Diagnóstico: No hay fractura.



Diagnóstico: **Fractura Schatzker tipo IV**.

Figure 2. We observed an error made by the AI, after several images were analysed. It should be noted that AI can have episodes of collapse or malfunction. Gr.2.

Diagnóstico: no hay fractura	Diagnosis: no fracture
Diagnóstico: fractura Schatzker tipo IV	Diagnosis: Schatzker type IV fracture



Diagnóstico: **No hay fractura.**

Las ocho imágenes muestran radiografías anteroposteriores de la rodilla sin evidencia de fractura en la meseta tibial. No se observan líneas de fractura, hundimientos ni desplazamientos óseos significativos.

Figure 3. Diagnostic capacity of multiple images. Errors were observed with more than 10 images. Gr.3.

Imagen	Image
Diagnóstico: no hay fractura Las ocho imágenes muestran radiografías anteroposteriores de la rodilla sin evidencia de fractura en la meseta tibial. No se observan líneas de fractura, hundimientos ni desplazamientos óseos significativos	Diagnosis: no fracture The eight images show anteroposterior radiographs of the knee without evidence of a tibial plateau fracture. No fracture lines, depressions, or significant bone displacements were observed.

Table 1. Comparison of the absolute agreement and Kappa value regarding fracture diagnosis

	[0,2-3]AI		Absolute agreement (%)	Kappa	[0,6-7]95%CI		p
Fracture	Absence	Presence			Inferior	Superior	
Absence	256	2	99.5	.988	.972	1.000	<.001
Presence	0	129					

	[0,2-3]Residents		Absolute agreement (%)	Kappa	[0,6-7]95%CI		P
Fracture	Absence	Presence			Inferior	Superior	
Absence	710	64	93.0	.848	.816	.880	<.,001
Presence	17	370					

	[0,2-3]Specialists		Absolute agreement (%)	Kappa	[0,6-7]95%CI		p
Fracture	Absence	Presence			Inferior	Superior	
Absence	745	17	97.0	.932	.910	.955	<.001
Presence	17	356					

AI: Artificial intelligence; CI: confidence interval.

Table 2. Comparison of the absolute agreement and the Kappa value regarding fracture classification

Evaluator	Absolute agreement (%)	Kappa	95%CI inferior	95%CI superior	p
AI	84.5	.612	.502	.722	<.001
Residents	82.8	.572	.510	.635	<.001
Specialists	84.8	.616	.554	.679	<.001

AI: Artificial intelligence; CI: confidence interval.

Table 3. Error Record by Schatzker Classification Grade

Schatzker Grade	[0,2-7]AI Grade					
	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 Grade	[0,2-7]Residents Grade					
	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 Grade	[0,2-7]Specialists Grade					
	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

The AI Grade table is corrected × 3. The tables for attending physicians and residents compile the sum of the errors of the 3 members of each group. It is important to indicate the accuracy of AI in type VI, unlike types I and IV.

AI: artificial intelligence.