Emergency Radiology (2025) 32:59–64
https://doi.org/10.1007/s10140-024-02298-y

RESEARCH

Diagnostic performance of ChatGPT in tibial plateau fracture in knee X-ray

Mohammadreza Mohammadi[1] · Sara Parviz[2] · Parinaz Parvaz[3] · Mohammad Mahdi Pirmoradi[1] · Mohammad Afzalimoghaddam[1,4] · Hadi Mirfazaelian[4]

Received: 13 September 2024 / Accepted: 30 October 2024 / Published online: 30 November 2024

Abstract

Purpose Tibial plateau fractures are relatively common and require accurate diagnosis. Chat Generative Pre-Trained Transformer (ChatGPT) has emerged as a tool to improve medical diagnosis. This study aims to investigate the accuracy of this tool in diagnosing tibial plateau fractures.

Methods A secondary analysis was performed on 111 knee radiographs from emergency department patients, with 29 confirmed fractures by computed tomography (CT) imaging. The X-rays were reviewed by a board-certified emergency physician (EP) and radiologist and then analyzed by ChatGPT-4 and ChatGPT-4o. The diagnostic performances were compared using the area under the receiver operating characteristic curve (AUC). Sensitivity, specificity, and likelihood ratios were also calculated.

Results The results indicated a sensitivity and negative likelihood ratio of 58.6% (95% CI: 38.9 −76.4%) and 0.4 (95% CI: 0.3–0.7) for the EP, 72.4% (95% CI: 52.7 −87.2%) and 0.3 (95% CI: 0.2–0.6) for the radiologist, 27.5% (95% CI: 12.7 −47.2%) and 0.7 (95% CI: 0.6–0.9) for ChatGPT-4, and 55.1% (95% CI: 35.6 −73.5%) and 0.4 (95% CI: 0.3–0.7) for ChatGPT4o. The specificity and positive likelihood ratio were 85.3% (95% CI: 75.8 −92.2%) and 4.0 (95% CI: 2.1–7.3) for the EP, 76.8% (95% CI: 66.2 −85.4%) and 3.1 (95% CI: 1.9–4.9) for the radiologist, 95.1% (95% CI: 87.9 −98.6%) and 5.6 (95% CI: 1.8–17.3) for ChatGPT-4, and 93.9% (95% CI: 86.3 −97.9%) and 9.0 (95% CI: 3.6–22.4) for ChatGPT4o. The area under the receiver operating characteristic curve (AUC) was 0.72 (95% CI: 0.6–0.8) for the EP, 0.75 (95% CI: 0.6–0.8) for the radiologist, 0.61 (95% CI: 0.4–0.7) for ChatGPT-4, and 0.74 (95% CI: 0.6–0.8) for ChatGPT4-o. The EP and radiologist significantly outperformed ChatGPT-4 (P value = 0.02 and 0.01, respectively), whereas there was no significant difference between the EP, ChatGPT-4o, and radiologist.

Conclusion ChatGPT-4o matched the physicians' performance and also had the highest specificity. Similar to the physicians, ChatGPT chatbots were not suitable for ruling out the fracture.

Keywords Tibial plateau fracture · ChatGPT · Artificial intelligence · Emergency medicine · Radiology · Diagnosis

✉ Hadi Mirfazaelian
H-Mirfazaelian@sina.tums.ac.ir

Mohammadreza Mohammadi
mr-mohammadi@student.tums.ac.ir

Sara Parviz
sarapz4u@yahoo.com

Parinaz Parvaz
Parinazparvaz744@gmail.com

Mohammad Mahdi Pirmoradi
mmpirmoradi76@gmail.com

Mohammad Afzalimoghaddam
afzalimoghadam@tuma.ac.ir

[1] Emergency Medicine Department, Tehran University of Medical Sciences, Tehran, Iran

[2] Musculoskeletal Imaging Research Center (MIRC), Tehran University of Medical Sciences, Tehran, Iran

[3] Radiology Department, Tehran University of Medical Sciences, Tehran, Iran

[4] Prehospital and Hospital Emergency Research Center, Tehran University of Medical Sciences, Tehran, Iran

Background

The tibial plateau is one of the most important load-bearing areas in the human body. Its fractures account for nearly 1% of all adult fractures and up to 8% of fractures in the elderly [1]. In particular, their incidence has increased over the last decade [2], which causes significant morbidity and economic burden [3]. Accurate imaging and diagnosis are required in these patients because of their role in the treatment plan [4, 5]. At the same time, even with the presence of decision rules for screening patients [6], this condition is among the most common missed radiological abnormalities mainly due to misinterpretation of radiographs [7, 8].

In the field of medical diagnostics, artificial intelligence (AI) has emerged as a tool for improving the accuracy and efficiency of clinical decision-making. A notable advance in this area has been the introduction of generative language models such as Chat Generative Pre-Trained Transformer (ChatGPT) with deep learning capabilities. ChatGPT is an advanced language model developed by OpenAI. It was able to understand and generate human-like text based on deep learning techniques [9, 10]. In September 2023, the premium version (ChatGPT 4) was released to handle image input and processing. These visual capabilities, powered by convolutional neural networks (CNNs), were developed through a training process similar to that used for ChatGPT 4's text processing. This technology, when applied to medical imaging has shown the potential to aid the diagnosis process [11]. In addition, recently, OpenAI introduced GPT-4o, an advanced model that accepts text, audio, image, and video inputs and generates real-time outputs with more accuracy.

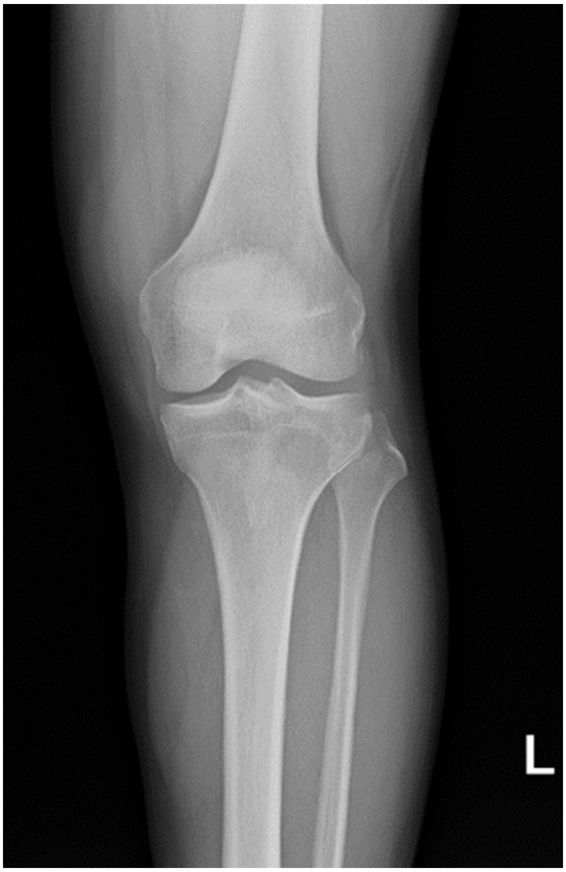


Fig. 1 A knee X-ray in anteroposterior (AP) view

The integration of ChatGPT into the interpretation of knee radiographs is expected to enhance the clinician's skills by leveraging the capabilities of AI. This study aims to explore the accuracy of ChatGPT in diagnosing tibial fractures and compare it with the emergency physician (EP) and the radiologist.

Methods

This study was conducted at a referral hospital affiliated with a major university in December 2023. The study was designed to evaluate the accuracy of the AI chatbot ChatGPT-4 and ChatGPT-4o for diagnosing tibial plateau fractures based on knee radiographs. This study has been reported according to standards for reporting diagnostic accuracy studies (STARD) reporting guidelines [12]. The study was approved by the institutional ethical review board at the university (IR.TUMS.IKHC.REC.1403.323) and conducted following the Declaration of Helsinki.

Design and setting

This was a secondary analysis of an unpublished study. It was a retrospective observational study on the imaging of a random sample of ED patients who underwent both knee radiography and computed tomography (CT) scans for acute knee trauma (either isolated or multiple trauma). The imaging had been ordered at the discretion of emergency physicians or trauma team leaders. The data of 111 cases was collected during 2023. Among them, 29 (26.13%) and 82 (73.87%) were with and without tibial plateau fractures, respectively. Images were in anteroposterior view, JPG format, with a horizontal and vertical resolution of 96 dpi. All the images were irreversibly anonymized and were presented to ChatGPT and the physicians. An example of the images is provided in Fig. 1 and Supplementary Information.

Study protocol and measurements

The images were interpreted by a board-certified EP and a radiologist both with more than 10 years of experience who did not know the CT scan findings. During X-ray image interpretation, they were only allowed to change the magnification of the images without any other modification. The images were uploaded to the ChatGPT-4 and ChatGPT-4o and two questions were asked consecutively:

- The following image shows the AP view of a knee X-ray. Can you detect a fracture in the image? Yes or No.
- If the answer was yes – Which bone is broken in the uploaded image?

After 2 weeks, the same images were uploaded to ChatGPT-4 in a different order and the same questions were asked to assess the reliability of the chatbot.

Outcome

The outcome of interest was the diagnostic accuracy of ChatGPT-4 and ChatGPT-4o in comparison to the EP and radiologist. The definite diagnosis was determined for each imaging based on its spiral knee CT scan interoperation by two other radiologists. During the CT image interpretation, the radiologists were able to reconstruct the images and make any modifications.

Statistical analysis

Normally distributed continuous variables were presented as mean with standard deviation and categorical variables were presented as numbers (percentage). Since this was a secondary analysis of another study, no sample size calculation was performed.

Multivariable logistic regression analysis was incorporated to compare the physicians' and ChatGPT's diagnostic performance. The intra-rater agreement (Kappa Statistic) test was also performed to evaluate the reliability of ChatGPT over a 2-week interval. The area under the receiver operating characteristic curve (AUC) was used to compare the proportion of discordant patient pairs (where one member of the pair had the fracture and the other did not), in which the patient with the fracture had the higher predicted probability. AUCs were compared according to the Delong et al. method [13]. In addition, 2 by 2 contingency tables were created to calculate sensitivity, specificity, and positive and negative likelihood ratio (+LR and –LR) values. The estimates and 95% confidence interval (CI) were reported. Analysis was conducted using RStudio 3.03.1 software. A p value < 0.05 was considered statistically significant.

Results

The average age of the study participants was 44 ± 15 years and 72% of the population were men. The diagnostic performance of different physicians and ChatGPTs was assessed. The results indicated a sensitivity of 58.6% (95% CI: 38.9 − 76.4%) for the EP, 72.4% (95% CI: 52.7 − 87.2%) for the radiologist, 27.5% (95% CI: 12.7-47.2%) for ChatGPT-4, and 55.1% (95% CI: 35.6 − 73.5%) for ChatGPT4o. The specificity was 85.3% (95% CI: 75.8 − 92.2%) for the EP, 76.8% (95% CI: 66.2 − 85.4%) for the radiologist, 95.1% (95% CI: 87.9 − 98.6%) for ChatGPT-4, and 93.9% (95% CI: 86.3 − 97.9%) for ChatGPT4o. +LR and -LR are presented in Table 1.

The discrimination ability of the physicians and the ChatGPTs was assessed using the receiver operating characteristic curve (Fig. 2). The area under the receiver operating characteristic curve (AUC) was 0.72 (95% CI: 0.6–0.8) for the EP, 0.75 (95% CI: 0.6–0.8) for the radiologist, 0.61(95% CI: 0.4–0.7) for ChatGPT-4, and 0.74 (95% CI: 0.6–0.8) for ChatGPT4-o (Table 1). The analysis showed that the EP and the radiologist outperformed the ChatGPT-4 significantly (P value = 0.02 and 0.01, respectively). The physicians' AUC was not significantly different from ChatGPT-4o. Of note, there was no significant difference between the radiologist and the EP either. (Table 2). The intra-rater agreement test also yielded a kappa value of 0.81, indicating a very good level of agreement in the responses of the ChatGPT over 2 weeks [14].

Discussion

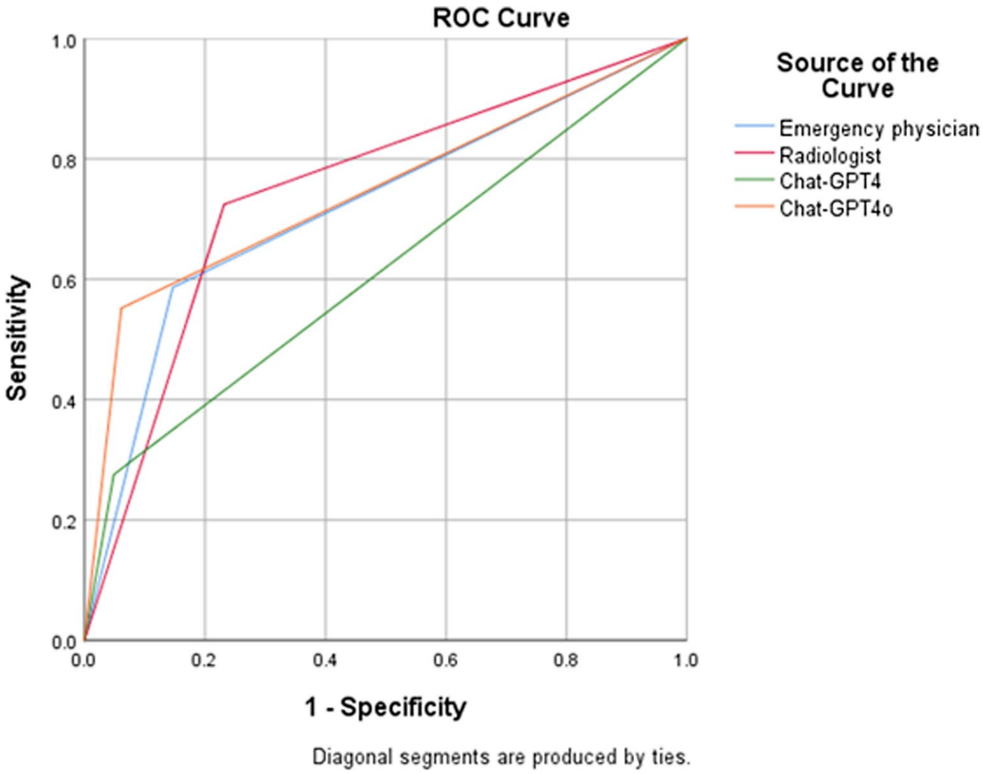
This study compared the diagnostic accuracy of the chatGPT-4 and chatGPT-4o chatbots with that of a radiologist and an EP. The radiologist showed the highest sensitivity and the ChatGPT-4o showed the highest specificity. While both of the physicians outperformed the ChatGPT, ChatGPT-4o performance reached the physicians' diagnosis discrimination ability. Furthermore, while ChatGPT-4o had the

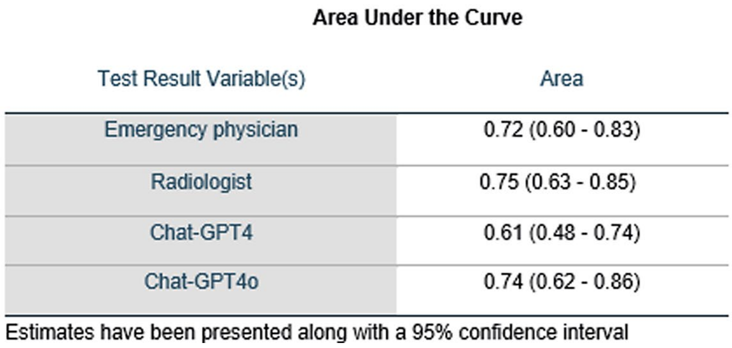
Table 1 Diagnostic performance indices of Emergency Medicine Physician, Radiologist, ChatGPT4, and ChatGPT-4o

	Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio	AUC
Emergency physician	58.6% (38.9 -76.4%)	85.3% (75.8 − 92.2%)	4.0 (2.1–7.3)	0.4 (0.3–0.7)	0.72 (0.60–0.83)
Radiologist	72.4% (52.7 − 87.2%)	76.8% (66.2 − 85.4%)	3.1(1.9–4.9)	0.3 (0.2–0.6)	0.75 (0.63–0.85)
ChatGPT-4	27.5% (12.7 − 47.2%)	95.1% (87.9 − 98.6%)	5.6 (1.8–17.3)	0.7 (0.6–0.9)	0.61(0.48–0.74)
ChatGPT-4o	55.1% (35.6 − 73.5%)	93.9% (86.3 − 97.9%)	9.0 (3.6– 22.4)	0.4 (0.3–0.7)	0.74 (0.62–0.86)

AUC: Area under the receiver operating characteristic curve, Estimates have been presented along with a 95% confidence interval

Fig. 2 ROC curve of the tibial plateau fracture detection





Area Under the Curve

Test Result Variable(s)	Area
Emergency physician	0.72 (0.60 - 0.83)
Radiologist	0.75 (0.63 - 0.85)
Chat-GPT4	0.61 (0.48 - 0.74)
Chat-GPT4o	0.74 (0.62 - 0.86)

Estimates have been presented along with a 95% confidence interval

Table 2 Comparison of the area under the receiver operating characteristic curve (AUC) of ChatGPT4, ChatGPT-4o, emergency medicine physician, and Radiologist

	Difference in AUC (95% confidence interval)	*P* value*
Radiologist - ChatGPT-4o	0.01 (-0.09–0.09)	0.98
Emergency physician - ChatGPT-4o	0.02 (-0.12–0.07)	0.10
Radiologist - ChatGPT-4	0.14 (0.01–0.24)	0.02
Emergency physician - ChatGPT-4	0.11 (-0.01–0.2)	0.01
Radiologist - Emergency physician	0.03 (-0.05–0.1)	0.53

*: Calculated using the De Long method

highest + LR and the lowest –LR, it cannot be used to rule out the fracture alone.

Various studies have evaluated different aspects of the ChatGPT chatbot application in medicine. For example, ChatGPT can provide context, background information, and differential diagnosis of diseases based on patient history, symptoms, and other diagnostic tests, making it a valuable tool in medical diagnostics [15]. In a recent study by Hirosawa et al., the diagnostic accuracy of differential diagnoses generated by an earlier version of ChatGPT was reported to be comparable to that of physicians [16]. Also, its application in radiology has been explored. For instance, ChatGPT may be integrated into radiology procedures to improve efficiency, accuracy, and patient care [17].

Previous studies have analyzed medical X-ray images using AI tools. Bousson V et al. evaluated the diagnostic accuracy of 3 available commercial AI algorithms designed for acute fracture detection on 1500 radiographs from 1210 consecutive patients with acute skeletal trauma. The accuracy rate was as high as 90.1% and the sensitivity was over 91%. They suggested that the body region is also important in this process apart from the AI algorithm. Interestingly, the highest accuracy rate was observed in the knee and the lowest was in the foot [18].

Liu PR et al. collected a dataset of 542 radiographs of tibial plateau fractures to build and train an AI algorithm. They compared the performance of the algorithm with 2 senior orthopedic surgeons in terms of accuracy and time spent on image analysis. Their AI algorithm showed an acceptable diagnostic accuracy of 0.91, similar to that of the physicians in the study. In addition, the AI algorithm significantly reduced the average time to diagnosis (16 times faster) compared to the orthopedic surgeons, which certainly needs further study as it could be a powerful adjunct in the clinic [19]. Also, the AI accuracy in their study surpasses both ChatGPT-4 and ChatGPT-4o in the current research. While this could be due to the better performance of the AI algorithm, it could also be due to the fact that the lateral views were not used in our study. The use of more views may improve diagnostic performance, as Gray et al. demonstrated that the use of 4 plain film views (by adding two oblique views to the standard views) increased the diagnostic sensitivity of knee fractures by 6% (from 79% to 85%) [20]. Another reason behind different findings might be the different study patient populations. Considering that our cases had CT imaging in addtion to the radiographs, we believe that our cases were from the population that were more difficult to identify the fracture hence needed further imaging (i.e., CT imaging).

Only one study evaluated the diagnostic performance of ChatGPT-4 in fractures. They used 150 images of wrist radiographs and compared them with those of a hand surgery resident, a medical student, and an AI application (BoneView) for distal radius fracture detection. They showed that ChatGPT-4 had good diagnostic performance (sensitivity 0.88, specificity 0.98, and AUC 0.93). It outperformed the medical student but fell short of a hand surgery resident and the AI application. Similarly, our study found that while ChatGPT-4 showed promising diagnostic capabilities, it was surpassed by the radiologist and the EP. Interestingly, the ChatGPT variant, the ChatGPT-4o likelihood ratios, and the AUC were similar to humans. A notable difference between the studies lies in the contexts. Mert et al. study focused on distal radius fractures, whereas current research assessed the knee with different anatomical complexity. In addition, having only one view of the diagnosis might play a role. Another potential explanation might be the training data differences highlighted by Mert et al. [11].

From the clinical perspective, accurate diagnosis of the plateau fracture is essential for effective management and to reduce the risk of complications. Although there are decision rules with acceptable miss rate to guide EPs with ordering radiographs [6], this process can be challenging due to ambiguous radiographic presentations (e.g., nondisplaced or anterior tibial fractures) [19, 21]. As shown in our study, image interpretation is not sensitive enough even in the hands of expert physicians. In addition to ChatGPT's ease of application, considering the performance indices such as modest sensitivity and high specificity of ChatGPT, it cannot be used to rule out the diagnosis but can serve as an adjunct to clinical decision-making.

Limitations

First, this study was based only on the patients' radiographs, without taking into account their history or physical examination. Furthermore, as mentioned above, since this was a secondary analysis, a lateral view of the knee X-rays was not available. As a result, important clinical parameters that may affect decision-making may have been absent. Also, the effect of data training on the diagnostic performance of ChatGPT was not assessed since the database used to train ChatGPT and its quality are unclear. Finally, we did not assess the time to interpretation of the images in our study. This process is probably faster for the AI and can spare more time for clinical practice for physicians.

Conclusion

In conclusion, the ChatGPT-4o performance was not different from the physicians' and even had the highest specificity. Of note, currently, neither the physicians nor the chatbots cannot be recommended to rule out the knee fracture alone.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10140-024-02298-y.

Author contributions MM: Writing – original draft, Writing – review & editing, Methodology, Data curation, Formal analysis. SP: Conceptualization, Supervision. PP: Data curation and validation. MMP: Conceptualization. MA: Validation, Supervision. HM: Conceptualization, Validation, Writing – original draft. All authors reviewed the manuscript.

Funding No funding or grants were received for performing this research.

Data availability De-identified case data are available from the corresponding author(H-Mirfazaelian@sina.tums.ac.ir) upon reasonable request.H-Mirfazaelian@sina.tums.ac.ir.

Declarations

Ethics approval and consent to participate The study was approved by the institutional ethical review board at the university and conducted following the Declaration of Helsinki.

Competing interests The authors declare no competing interests.

References

1. Herteleer M, Van Brandt C, Vandoren C, Nijs S, Hoekstra H (2022) Tibial plateau fractures in Belgium: epidemiology, financial burden and costs curbing strategies. Eur J Trauma Emerg Surg 48(5):3643–3650
2. Bormann M, Neidlein C, Gassner C, Keppler AM, Bogner-Flatz V, Ehrnthaller C et al (2023) Changing patterns in the epidemiology of tibial plateau fractures: a 10-year review at a level-I trauma center. Eur J Trauma Emerg Surg 49(1):401–409
3. Ramponi DR, McSwigan T (2018) Tibial Plateau fractures. Adv Emerg Nurs J 40(3):155–161
4. Rudran B, Little C, Wiik A, Logishetty K (2020) Tibial Plateau fracture: anatomy, diagnosis and management. Br J Hosp Med (Lond) 81(10):1–9
5. Schatzker J, Kfuri M (2022) Revisiting the management of tibial plateau fractures. Injury 53(6):2207–2218
6. Stiell IG, Greenberg GH, Wells GA, McKnight RD, Cwinn AA, Cacciotti T et al (1995) Derivation of a decision rule for the use of radiography in acute knee injuries. Ann Emerg Med 26(4):405–413
7. Kiel CM, Mikkelsen KL, Krogsgaard MR (2018) Why tibial plateau fractures are overlooked. BMC Musculoskelet Disord 19(1):244
8. Sprivulis P, Frazer A, Waring A (2001) Same-day X-ray reporting is not needed in well-supervised emergency departments. Emerg Med (Fremantle) 13(2):194–197
9. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I et al GPT-4 Technical Report2023 March 01, 2023:[arXiv:2303.08774 p.]. https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O
10. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S et al (2024) Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 66(1):73–79
11. Mert S, Stoerzer P, Brauer J, Fuchs B, Haas-Lützenberger EM, Demmer W et al (2024) Diagnostic power of ChatGPT 4 in distal radius fracture detection through wrist radiographs. Arch Orthop Trauma Surg 144(5):2461–2467
12. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L et al (2016) STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6(11):e012799
13. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–845
14. Byrt T (1996) How good is that agreement? Epidemiology 7(5):561
15. Tustumi F, Andreollo NA, Aguilar-Nascimento JE, FUTURE OF THE LANGUAGE, MODELS IN HEALTHCARE: THE ROLE OF CHATGPT (2023) Arq Bras Cir Dig 36:e1727
16. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T (2023) Diagnostic accuracy of Differential-diagnosis lists generated by Generative Pretrained Transformer 3 Chatbot for Clinical vignettes with Common Chief complaints: a pilot study. Int J Environ Res Public Health.;20(4)
17. Mese I, Taslicay CA, Sivrioglu AK (2023) Improving radiology workflow using ChatGPT and artificial intelligence. Clin Imaging 103:109993
18. Bousson V, Attané G, Benoist N, Perronne L, Diallo A, Hadid-Beurrier L et al (2023) Artificial Intelligence for detecting Acute fractures in patients admitted to an Emergency Department: real-life performance of three commercial algorithms. Acad Radiol 30(10):2118–2139
19. Liu PR, Zhang JY, Xue MD, Duan YY, Hu JL, Liu SX et al (2021) Artificial Intelligence to diagnose Tibial Plateau fractures: an Intelligent Assistant for Orthopedic Physicians. Curr Med Sci 41(6):1158–1164
20. Gray SD, Kaplan PA, Dussault RG, Omary RA, Campbell SE, Chrisman HB et al (1997) Acute knee trauma: how many plain film views are necessary for the initial examination? Skeletal Radiol 26(5):298–302
21. Maheshwari J, Pandey VK, Mhaskar VA (2014) Anterior tibial plateau fracture: an often missed injury. Indian J Orthop 48(5):507–510

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.