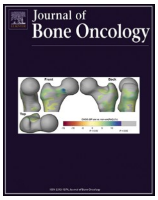
J BONE ONCOL 48 (2024) 100632

Contents lists available at ScienceDirect

Journal of Bone Oncology

journal homepage: www.elsevier.com/locate/jbo



Short Communication

Preliminary discrimination and evaluation of clinical application value of ChatGPT4o in bone tumors

Leiyun Huang [a,b,c,d,e,1], Jinghan Hu [f,1], Qingjin Cai [g,1], Aoran Ye [h], Yanxiong Chen [a,b,c,d,e], Zha Yang Xiao-zhi [a,b,c,d,e], Yongzhen Liu [f], Ji Zheng [g], Zengdong Meng [a,b,c,d,e,*]

[a] *Medical College, Kunming University of Science and Technology, Kunming, China*
[b] *Department of Orthopedics, The First People's Hospital of Yunnan Province, Kunming, China*
[c] *Key Laboratory of Digital Orthopedics of Yunnan Province, Kunming, China*
[d] *Clinical Medicine Research Center of Orthopedics and Sports Rehabilitation in Yunnan Province, Kunming, China*
[e] *Clinical Medical Center for Spinal Cord Diseases in Yunnan Province, Kunming, China*
[f] *People's Hospital of Wenshan Prefecture, the Affiliated Hospital of Kunming University of Science and Technology, Wenshan, China*
[g] *Department of Urology, Urologic Surgery Center, Xinqiao Hospital, Third Military Medical University (Army Medical University), Chongqing 400037, China*
[h] *Department of Pain, The First People's Hospital of Yunnan Province, Kunming, China*

HIGHLIGHTS

- Evaluation of making up ChatGPT4o in the preliminary pathological diagnosis of bone tumors.
- ChatGPT-4o's proficiency in analyzing pathological images and providing initial diagnoses of bone tumor characteristics is comparable to that of senior pathologists in the Tertiary hospital doctors group, with both surpassing the Remote grassroots doctors group.
- AI, like ChatGPT-4o, has the potential to enhance diagnostic capabilities for remote grassroots doctors and improve sensitivity to reduce missed diagnosis rates among tertiary hospital doctors in identifying bone tumors.

ARTICLE INFO

Keywords:
ChatGPT4o
Clinical medicine
Pathology
Medical education

Dear Editor

1. Introduction

Radiographic imaging and pathological results are the two most critical methods for clinicians to distinguish between benign and malignant bone tumors. Fan Yang et al. [1] assessed the performance of ChatGPT3.5 in diagnosing benign and malignant bone tumors through imaging reports, highlighting the potential application of artificial intelligence (AI) technology in medical imaging diagnosis, particularly in enhancing diagnostic efficiency and reducing missed diagnoses. However, they did not evaluate the performance of ChatGPT in pathological images. To fill this gap, we evaluated the preliminary ability of ChatGPT4o in discerning bone tumors through pathological images and to optimize secondary prevention of malignant bone tumors in remote grassroots orthopedic settings. The study seeks to improve the diagnostic and treatment capabilities of remote grassroots doctors, assisting clinicians in early screening, diagnosis, and treatment. Additionally, it aims to ensure timely referral of patients to higher-level hospitals when constrained by objective conditions, thereby enhancing cancer cure rates and survival rates.

[*] Corresponding author at: Department of Orthopedics, The First People's Hospital of Yunnan Province, Kunming, China
E-mail address: menggu7119@vip.sina.com (Z. Meng).
[1] Contributed equally.

https://doi.org/10.1016/j.jbo.2024.100632
Received 17 July 2024; Received in revised form 22 August 2024; Accepted 27 August 2024
Available online 2 September 2024

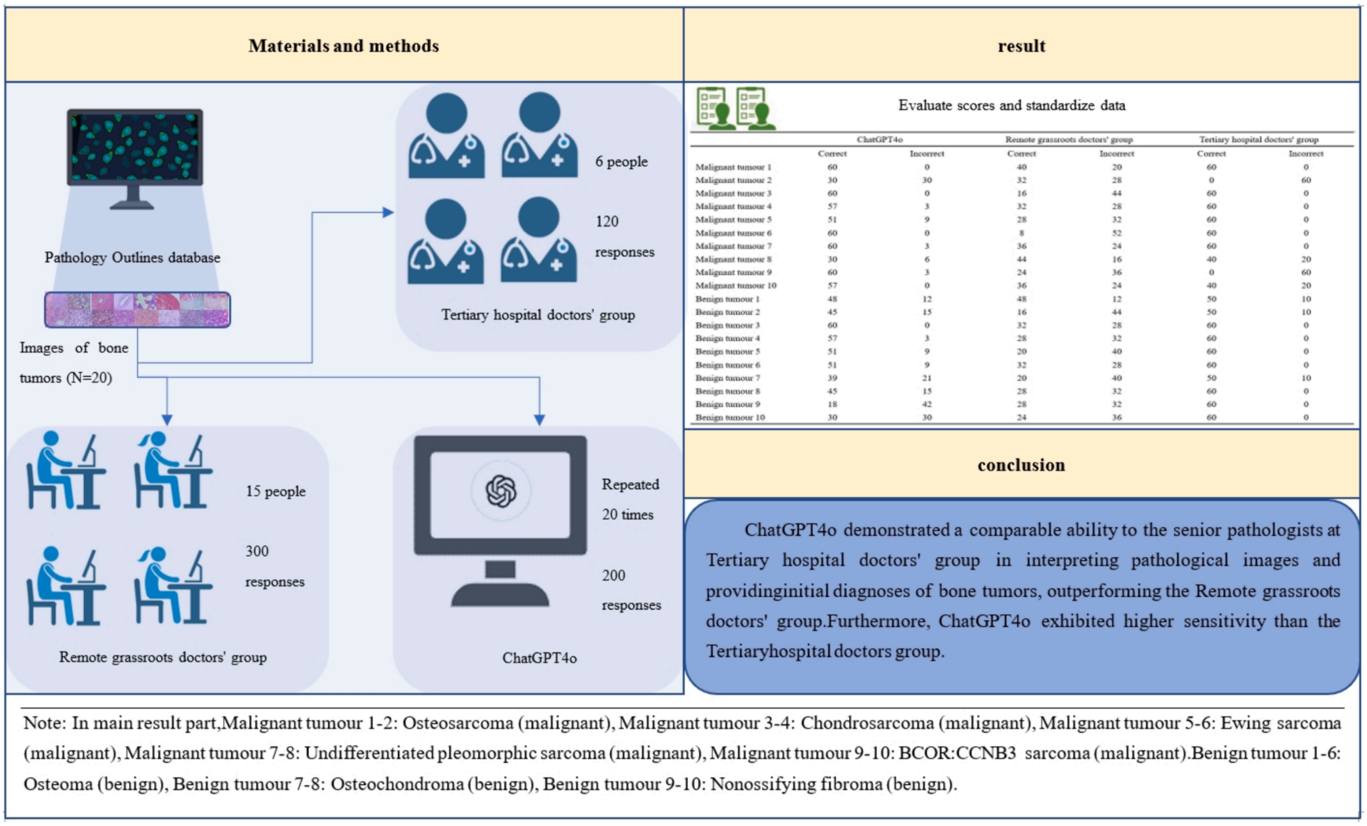
Twenty pathological images of bone tumors were randomly selected from the Pathology Outlines database, and all pathological images are displayed in the supplementary materials (Supplementary Material 1). The procedural workflow is outlined in the technical roadmap (Fig. 1). Among these images, two were used in a preliminary experiment, which revealed that this database is not within the “pre-training” scope of ChatGPT-4o (Supplementary Material 2). Similarly, the study by Yiwen Zhang et al. [2] indirectly supported this finding. Fifteen remote grassroots doctors (from township and community hospitals) of various professional titles, six senior pathologists from tertiary hospitals, and ChatGPT4o participated in this preliminary diagnostic test. No additional information was provided during the entire process, resulting in a total of 820 diagnostic feedbacks. Among these, 420 responses were diagnoses made by doctors, while 400 responses were obtained from various members of our research team at different locations and time points, all operating ChatGPT4o with the same prompts. To avoid cumbersome prompts, each image was independently presented to ChatGPT4o in a separate dialogue interface 20 times with the prompt: “As a senior expert in clinical medicine and pathology, can you describe this histological image of a bone/cartilage/fibrous tissue tumor using HE staining? Does this image represent a benign or malignant bone/cartilage/fibrous tissue tumor?” All responses were quantified as scores: “1 = correct, 0 = incorrect.” Additionally, the quantized scores were standardized to the lowest common multiple of 60 to derive the preliminary diagnostic performance of the pathological images (Fig. 2).

The study revealed significant differences in accuracy among the Remote grassroots doctors’ group (Rgdg), multimodal ChatGPT4o (GPT4o), and the Tertiary hospital doctors’ group (Thdg) (Kruskal-Wallis test, $p<0.001$). However, there was no significant variance between ChatGPT4o and the Tertiary hospital doctors’ group ($p = 0.957$). This indicates that ChatGPT4o’s proficiency in analyzing pathological images and providing initial diagnoses of bone tumor characteristics is comparable to that of senior pathologists in the Tertiary hospital doctors’ group, with both surpassing the Remote grassroots doctors’ group. Their diagnostic performance is as follows (Table 1). The reproducibility of the diagnoses presented in Figures 8 and 16 of ChatGPT4o may be limited. This observation indicates that ChatGPT4o may have constraints in diagnosing specific pathological images. In contrast, other images exhibit strong reproducibility, as shown in Fig. 2.

This study has certain limitations. For example, we have not yet conducted additional evaluations across various pathological databases, staining methods, or staining intensities. Future research could address these limitations by expanding the pathological databases, increasing the sample size and diversity of images, and including physicians with varying levels of experience.

Pathological biopsy is considered the gold standard for diagnosing both benign and malignant bone tumors. However, many doctors in remote grassroots settings lack the specialized training needed to accurately interpret pathological results. In such situations, the use of AI to assist doctors in determining the nature of bone tumors can provide significant clinical value, particularly in geographically isolated and medically underserved areas. The study conducted by Nigam H Shah et al. [3] highlights the increasing interest and potential advantages of utilizing large language models (LLMs) in the field of medicine. Applications driven by LLMs are increasingly being utilized for various medical tasks. Nonetheless, in order to validate the benefits of these models in actual task execution, further evaluation through real-world deployment tests is essential.

In the diagnosis of malignant tumors, an ideal diagnostic tool should have nearly 100% sensitivity and specificity. However, achieving this in actual clinical practice is challenging. Increasing sensitivity can lead to higher misdiagnosis rates, while increasing specificity can result in higher missed diagnosis rates [4,5]. To reduce the misdiagnosis rate and prevent unnecessary fear among patients, specificity must be improved. Conversely, to detect malignant tumors early and provide timely treatment, sensitivity must be enhanced to reduce missed diagnosis rates. AI,

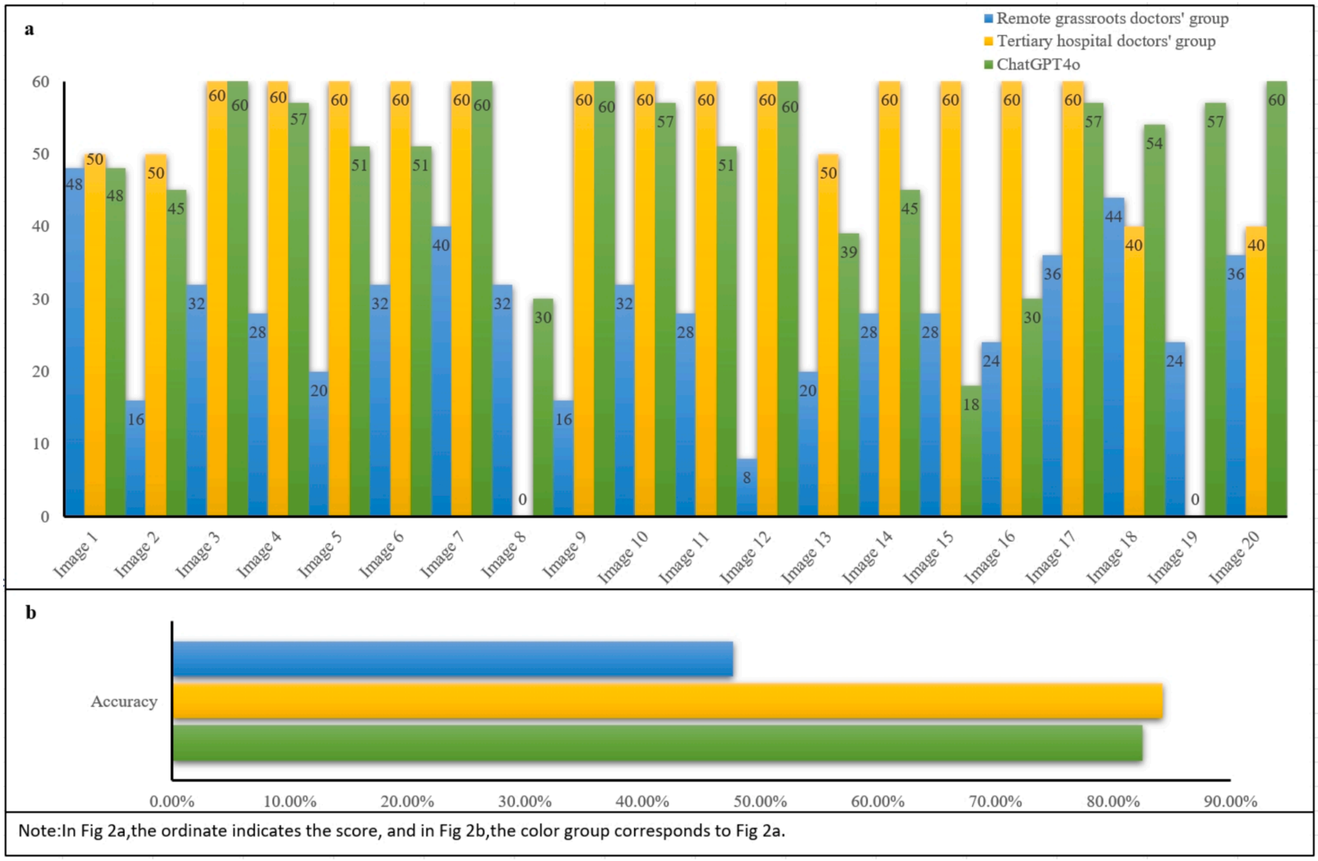

Materials and methods
Pathology Outlines database
Images of bone tumors (N=20)
6 people
120 responses
Tertiary hospital doctors' group
15 people
300 responses
Remote grassroots doctors' group
Repeated 20 times
200 responses
ChatGPT4o
result
Evaluate scores and standardize data
conclusion

	ChatGPT4o		Remote grassroots doctors' group		Tertiary hospital doctors' group	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Malignant tumour 1	60	0	40	20	60	0
Malignant tumour 2	30	30	32	28	0	60
Malignant tumour 3	60	0	16	44	60	0
Malignant tumour 4	57	3	32	28	60	0
Malignant tumour 5	51	9	28	32	60	0
Malignant tumour 6	60	0	8	52	60	0
Malignant tumour 7	60	3	36	24	60	0
Malignant tumour 8	30	6	44	16	40	20
Malignant tumour 9	60	3	24	36	0	60
Malignant tumour 10	57	0	36	24	40	20
Benign tumour 1	48	12	48	12	50	10
Benign tumour 2	45	15	16	44	50	10
Benign tumour 3	60	0	32	28	60	0
Benign tumour 4	57	3	28	32	60	0
Benign tumour 5	51	9	20	40	60	0
Benign tumour 6	51	9	32	28	60	0
Benign tumour 7	39	21	20	40	50	10
Benign tumour 8	45	15	28	32	60	0
Benign tumour 9	18	42	28	32	60	0
Benign tumour 10	30	30	24	36	60	0

ChatGPT4o demonstrated a comparable ability to the senior pathologists at Tertiary hospital doctors' group in interpreting pathological images and providinginitial diagnoses of bone tumors, outperforming the Remote grassroots doctors' group.Furthermore, ChatGPT4o exhibited higher sensitivity than the Tertiaryhospital doctors group.

Note: In main result part,Malignant tumour 1-2: Osteosarcoma (malignant), Malignant tumour 3-4: Chondrosarcoma (malignant), Malignant tumour 5-6: Ewing sarcoma (malignant), Malignant tumour 7-8: Undifferentiated pleomorphic sarcoma (malignant), Malignant tumour 9-10: BCOR:CCNB3 sarcoma (malignant).Benign tumour 1-6: Osteoma (benign), Benign tumour 7-8: Osteochondroma (benign), Benign tumour 9-10: Nonossifying fibroma (benign).

Fig. 1. Technical route and operation flow.



Note:In Fig 2a,the ordinate indicates the score, and in Fig 2b,the color group corresponds to Fig 2a.

Fig. 2. Distribution of score frequency after standardization.

Table 1

Diagnostic performance	Rgdg	Thdg	GPT4o
True Positive	296	440	525
False Positive	324	30	156
True Negative	276	570	444
False Negative	304	160	54
Sensitivity	0.49	0.73	0.91
Specificity	0.46	0.95	0.74
Misdiagnosis Rate	0.54	0.05	0.26
Missed Diagnosis Rate	0.51	0.27	0.09

Note: In Table 1, 'True Positive' refers to cases where a malignant tumor is diagnosed by a doctor or AI, and it is indeed malignant. 'False Positive' refers to cases where a malignant tumor is diagnosed by a doctor or AI, but it is actually benign. 'True Negative' refers to cases where a benign tumor is diagnosed by a doctor or AI, and it is indeed benign. 'False Negative' refers to cases where a benign tumor is diagnosed by a doctor or AI, but it is actually malignant.“.

like ChatGPT4o, has the potential to enhance diagnostic capabilities for Remote grassroots doctors and improve sensitivity to reduce missed diagnosis rates among Tertiary hospital doctors in identifying bone tumors. The combination of doctors and Artificial Intelligence (AI) can help achieve a balance between sensitivity and specificity in diagnosis, minimizing misdiagnoses and patient panic while ensuring timely treatment. It is important to emphasize that the results of this study are preliminary and require further validation. Artificial intelligence tools should only be utilized by individuals with a medical background for the purposes of medical education and diagnostic support. Furthermore, the application of AI in healthcare should be governed by stringent regulations.

Funding

This research was supported by the Yunnan Provincial Health Commission's Leading Medical Talent Training Program (L-218004), the Yunnan Provincial Xingdian Elite Support Program for Outstanding Doctors (XDYC-MY-2022-0027), the Yunnan Provincial Clinical Medical Research Center for Orthopedics and Sports Rehabilitation (202102AA310068), the Yunnan Provincial Clinical Medical Center for Spinal Cord and Spinal Diseases (ZX2022000101), the Key Laboratory of Digital Orthopedics of Yunnan Province (202005AG070004), the Key Research and Development Program of Yunnan Provincial Science and Technology Plan (202403AC100003), and the Yunnan Key Laboratory of Digital Orthopedics Platform open project (2024SGKFKT-02).

Ethics approval and consent to participate

Not applicable.

CRediT authorship contribution statement

Leiyun Huang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Data curation. **Jinghan Hu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Data curation. **Qingjin Cai:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Data curation. **Aoran Ye:** Writing – original draft, Validation, Supervision, Investigation. **Yanxiong Chen:** Writing – original draft, Validation, Supervision, Investigation. **Zha Yang Xiao-zhi:** Writing – original draft, Writing – review & editing. **Yongzhen Liu:** Writing – original draft, Validation, Supervision. **Ji Zheng:** Writing – review & editing, Validation, Supervision. **Zengdong Meng:** Writing – review & editing, Writing

– original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbo.2024.100632.

References

[1] F. Yang, D. Yan, Z. Wang, Large-Scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications, J. Bone Oncol. 44 (2024), https://doi.org/10.1016/j.jbo.2024.100525, 100525.

[2] Y. Zhang, H. Liu, B. Sheng, Y.C. Tham, H. Ji, Preliminary fatty liver disease grading using general-purpose online large language models: ChatGPT-4 or Bard?, J. Hepatol. 80(6) (2024) e279-e281. Doi: 10.1016/j.jhep.2023.11.017.

[3] N.H. Shah, D. Entwistle, M.A. Pfeffer, Creation and Adoption of large language models in medicine, JAMA 330(9) (2023) 866-869. Doi: 10.1001/jama.2023.14217.

[4] K.J.L. Bell, J. Doust, P. Glasziou, L. Cullen, I.A. Harris, L. Smith, R. Buchbinder, A. Barratt, recognizing the potential for overdiagnosis: are high-sensitivity cardiac troponin assays an example?, Ann. Intern. Med. 170(4) (2019) 259-261. Doi: 10.7326/M18-2645.

[5] A.A. Tarnutzer, S.H. Lee, K.A. Robinson, Z. Wang, J.A. Edlow, D.E. Newman-Toker, ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: A meta-analysis, Neurology 88(15) (2017) 1468-1477 Doi: 10.1212/WNL.0000000000003814.