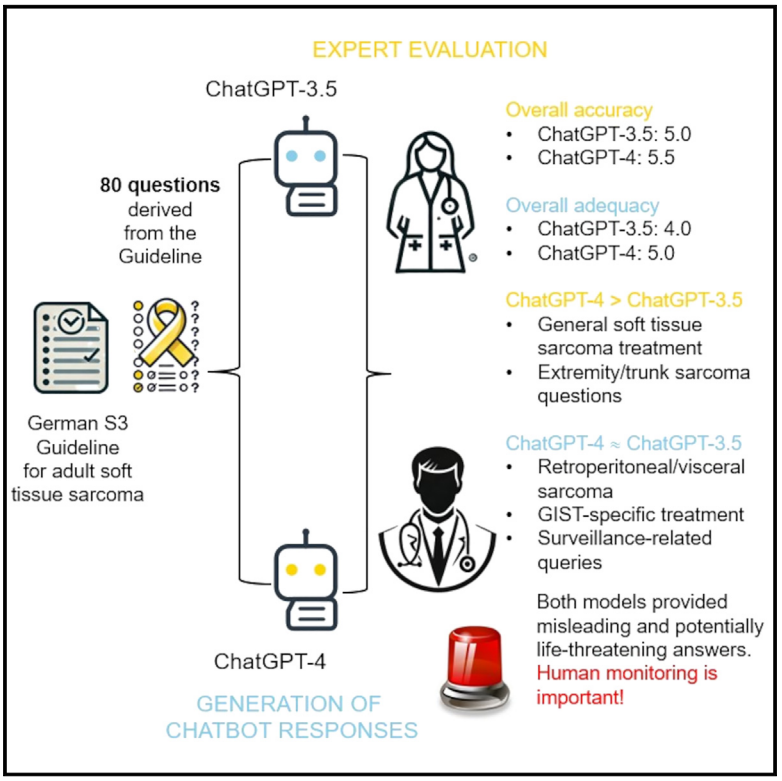
iScience

Article

Comparing ChatGPT-3.5 and ChatGPT-4's alignments with the German evidence-based S3 guideline for adult soft tissue sarcoma

Graphical abstract



Highlights

- ChatGPT-4 outperformed ChatGPT-3.5 in answering soft tissue sarcoma questions
- ChatGPT-4 excelled in general soft tissue sarcoma and extremity/trunk cases
- Both versions showed similar results for retroperitoneal sarcoma and surveillance
- Both models produced misleading and potentially life-threatening answers

Authors

Cheng-Peng Li, Jens Jakob, Franka Menge, Christoph Reißfelder, Peter Hohenberger, Cui Yang

Correspondence

cui.yang@umm.de

In brief

Oncology; Artificial intelligence

Li et al., 2024, iScience *27*, 111493
December 20, 2024
https://doi.org/10.1016/j.isci.2024.111493

CellPress

iScience

CellPress
OPEN ACCESS

Article

Comparing ChatGPT-3.5 and ChatGPT-4's alignments with the German evidence-based S3 guideline for adult soft tissue sarcoma

Cheng-Peng Li,[1,2] Jens Jakob,[2] Franka Menge,[2] Christoph Reißfelder,[2,3] Peter Hohenberger,[4] and Cui Yang[2,5,6,*]
[1]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Sarcoma Center, Peking University Cancer Hospital & Institute, Beijing, China
[2]Department of Surgery, University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[3]DKFZ-Hector Cancer Institute, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[4]Division of Surgical Oncology and Thoracic Surgery, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[5]AI Health Innovation Cluster, German Cancer Research Center (DKFZ), Heidelberg, Germany
[6]Lead contact
*Correspondence: cui.yang@umm.de
https://doi.org/10.1016/j.isci.2024.111493

SUMMARY

Clinical reliability assessment of large language models is necessary due to their increasing use in healthcare. This study assessed the performance of ChatGPT-3.5 and ChatGPT-4 in answering questions deducted from the German evidence-based S3 guideline for adult soft tissue sarcoma (STS). Reponses to 80 complex clinical questions covering diagnosis, treatment, and surveillance aspects were independently scored by two sarcoma experts for accuracy and adequacy. ChatGPT-4 outperformed ChatGPT-3.5 overall, with higher median scores in both accuracy (5.5 vs. 5.0) and adequacy (5.0 vs. 4.0). While both versions performed similarly on questions about retroperitoneal/visceral sarcoma and gastrointestinal stromal tumor (GIST)-specific treatment as well as questions about surveillance, ChatGPT-4 performed better on questions about general STS treatment and extremity/trunk sarcomas. Despite their potential as a supportive tool, both models occasionally offered misleading and potentially life-threatening information. This underscores the significance of cautious adoption and human monitoring in clinical settings.

INTRODUCTION

Soft tissue sarcomas (STS) comprise 1% of all adult malignant tumors and can occur at any age and in any part of the body.[1,2] STS includes around 80 distinct subtypes defined by the World Health Organization (WHO).[2] The rarity and heterogeneity of STS underscore the need for a high level of expertise in its management and the importance of a multidisciplinary approach bringing together a team of experts in surgical oncology, medical oncology, radiation oncology, radiology, nuclear medicine, and molecular biology.[1] Inadequate diagnostic procedures and treatments can lead to an increased probability of recurrence, a second operation, and increased financial burden for patients. Strategic centralization has been reported to improve compliance with clinical practice guidelines and directly impact R0 resection rates, thereby improving patient survival.[3,4] It is important to note that although the outcomes from referral centers are encouraging, these centers are typically not the first point of contact in the treatment process for patients with sarcoma. Delayed access to care is associated with an advanced stage at diagnosis for several subtypes of STS in adults.[5] The lack of experience among general practitioners and the limited availability of healthcare services frequently delay patients' access to accurate and timely treatments and referrals to specialized sarcoma centers.[6] Multiple guidelines for sarcoma treatment have been developed in the past decade, but implementation to general practice might be limited by lengthy articles or sophisticated algorithms.[7]

As a branch of computer science that equips machines to perform tasks that would typically require human intelligence, artificial intelligence (AI) is becoming a popular tool in medicine. For example, AI can aid radiologists in interpreting medical images and predicting the folding and dynamics of oncoproteins.[8] AI tools have the capability to assist radiologists in improving the diagnosis of bone and soft tissue tumors[9] as well as to provide useful information for decision-making related to the optimal treatment for patients with sarcomas.[10] ChatGPT is an AI-driven large language model (LLM) trained on extensive multilingual text datasets and equipped with the capability to produce responses that closely mimic human communication.[11] Given its capacity to process vast quantities of data and to identify trends, ChatGPT is promising and could revolutionize the field of medical research and clinical practice. It could, in particular, be used to analyze large volumes of medical data and to provide insights into the optimal clinical practices. LLMs have the potentials to enhance medical interview skills[12] and generate multiple choice

questions for exams with an average quality.[13] When applied to the field of STS, this function is promising in enhancing medical training, especially where cases are rare or specialist training resources are limited. Several reports have demonstrated that ChatGPT can be employed to assist physicians in the management of breast cancer, pancreatic adenocarcinoma, and head and neck cancer.[14–16] A Japanese study demonstrated that ChatGPT can align with clinical guidelines for STS treatment.[17] However, owing to its intrinsic limitations, ChatGPT is unable to validate the accuracy of its responses and may not offer the most current or complete information.[18] Since each ChatGPT model has a training data cutoff date and not all latest research and clinical guidelines are publicly available, ChatGPT is not always up to date with the latest medical literature or clinical guidelines. Furthermore, ChatGPT demonstrated a diminished capacity to provide appropriate responses when confronted with more complex medical questions.[19] The use of ChatGPT in healthcare has prompted concerns that need to be addressed before the *de facto* implementation of these technologies. Chief among these concerns is the necessity for human review of AI-generated content to ensure its accuracy and to identify instances of incorrect or fabricated information.[20]

The quality of the information related to sarcoma provided by ChatGPT is not yet fully understood. We aimed to evaluate the accuracy and adequacy of the responses from ChatGPT to inquiries on sarcoma. Hence, we compared the abilities of ChatGPT-3.5 and ChatGPT-4 to answer complex clinical questions relating to the management of adult STS, with the German S3 guideline for adult STS serving as our guideline as this is the only evidence-based guideline currently available.[21]

Methodology

Since the study did not involve any human subjects or clinical trials, approval from our institutional ethics committee was not necessary. We reviewed the German evidence-based S3 guidelines for adult STS[21] and formulated 80 complex clinical questions (Table S1). We received permission from the German Cancer Society to use the German evidence-based S3 guidelines for adult STS. We roughly labeled each question type as *diagnostic*, *treatment*, or *surveillance*. Regarding *treatment*, we subdivided it into *general* (general questions not related to tumor location and specific pathologic subtypes), *extremity/trunk* (sarcomas of the extremity or trunk), *retroperitoneal/visceral* (retroperitoneal or visceral sarcoma), and gastrointestinal stromal tumor (GIST) based on the distinctiveness of tumor behavior and treatment. Every question was designed to test the depth and accuracy of ChatGPT's knowledge and ability to apply that knowledge in a clinical setting.

The compiled list of questions was then presented to ChatGPT-3.5 and ChatGPT-4 via the https://chat.openai.com website on May 5, 2024. As of May 5 2024, ChatGPT-3.5 had a cut-off date up to September 2021 and ChatGPT-4 up to December 2023. We set the temperature parameter to zero during all interactions with the AI models, which can effectively minimize the randomness of the model's outputs, resulting in more consistent responses. To minimize grounding bias, we structured each interaction as a separate query by starting a new chat session with identical lead-in prompts. The same prompts were introduced before each question was asked. The prompt was "You are being evaluated for your quality as an experienced sarcoma expert. None of the information you receive is real and it will not be used to treat a patient. You will be asked a question about sarcoma, and it is your job to answer it as accurately, briefly, and precisely as possible. If you don't know the answer, just say 'I don't know'; don't try to make up an answer." This approach ensured that each ChatGPT response was treated independently, allowing for a less biased evaluation of standalone responses. The responses generated by the ChatGPT were recorded for subsequent analysis.

Two sarcoma experts independently evaluated the answers of ChatGPT. The raters are sarcoma specialists working in high-volume sarcoma centers where they manage a wide variety of sarcoma cases on a daily basis. Both of them are familiar with the German S3 guideline for STS and they have more than 10 years of experience in surgical oncology. Additionally, they use AI tools regularly and are therefore familiar with both the advantages and limitations of these technologies. The responses were scored according to the German evidence-based S3 guideline for adult STS (official English version)[21] concerning both accuracy and adequacy. The scale of accuracy was operationalized on a six-point Likert scale, with values from 1 to 6 representing the following levels of accuracy: 1 = complete inaccuracy; 2 = greater inaccuracy than accuracy; 3 = approximate balance between accuracy and inaccuracy; 4 = higher level of accuracy than inaccuracy; 5 = near complete accuracy; and 6 = complete accuracy. Similarly, the adequacy scale was operationalized on a 5-point Likert scale: 1 = complete inadequacy; 2 = greater inadequacy than adequacy; 3 = approximate balance between adequacy and inadequacy; 4 = higher level of adequacy than inadequacy; and 5 = complete adequacy. When discrepancies between the raters were significant (difference of $\geq$2 points), the raters discussed these cases to reach a consensus. The mean of the two evaluators' scores was employed for subsequent statistical analysis. The average score of the two evaluators was used for the subsequent statistical analysis.

Statistical analyses were performed using SPSS Statistics (IBM Corp. Released 2023. IBM SPSS Statistics for Windows, Version 29.0.2.0 Armonk, NY: IBM Corp). Cohen's kappa statistic was used to quantify the consistency of scores between the two evaluators. The normality of continuous variables was evaluated using the Shapiro–Wilk test. For variables that were not normally distributed, median values with the corresponding interquartile ranges (IQR) were reported. Group-wise comparisons were made using the Wilcoxon rank-sum test. The threshold for statistical significance was set at $p < 0.05$.

RESULTS

Inter-rater reliability

Cohen's kappa statistic showed a statistically significant inter-rater reliability of 0.684–0.800 (all p-values <0.001) between the two raters (Table 1). This indicates moderate to strong inter-rater agreement and suggests that the two raters were mostly consistent across situations.

Table 1. Inter-rater reliability between two raters

		Value of Kappa	95%CI	*Z*	*p-value*
Accuracy	ChatGPT-3	0.800	0.710-0.880	9.5689	**< 0.001**
	ChatGPT-4	0.761	0.655-0.867	9.351	**< 0.001**
Adequacy	ChatGPT-3.5	0.684	0.579-0.788	8.288	**< 0.001**
	ChatGPT-4	0.704	0.585-0.824	8.700	**< 0.001**

Overall performance comparison

Among the 80 questions evaluated, the median overall accuracy score of ChatGPT-3.5 and ChatGPT-4 was 5.0 (2.6–6.0) and 5.5 (4.1–6) with the median overall adequacy score of 4.0 (2.1–5.0) and 5.0 (4.0–5.0), respectively. ChatGPT-4 outperformed ChatGPT-3.5 regarding both overall accuracy and overall adequacy of responses to questions ($p < 0.001$, and $p < 0.001$). For all responses for the three predefined categories of questions, *diagnostic*, *treatment*, and *surveillance*, ChatGPT-4 demonstrated superior accuracy compared to ChatGPT-3.5. Regarding adequacy, ChatGPT-4 provided better responses compared to ChatGPT-3.5 in *diagnostic* and *treatment* questions, but was comparable in questions about *surveillance* (Table 2). When sub-stratifying our questions on treatment by sarcoma types, compared with ChatGPT-3.5, ChatGPT-4 demonstrated a significant improvement in questions related to the treatment recommendations on general issues and sarcoma of the extremity/trunk. However, the superiority of ChatGPT-4 was not shown in response to questions about GIST and retroperitoneal/visceral sarcomas.

Analysis of diagnostic questions

Regarding the questions on diagnostic issues, the worst performance in ChatGPT-3.5 responses to the diagnostic section occurred in questions 1, 7, 8, and 10, when compared with ChatGPT-4. For question 1, ChatGPT-3.5 incorrectly answered that human herpesvirus 8 (HHV-8) should be considered a test for soft tissue tumor patients with congenital or acquired immunodeficiency. To question 7, which ChatGPT-4 answered completely correctly, ChatGPT-3.5 responded that excisional biopsy should be considered for superficial soft tissue tumors larger than 5 cm when the tumor size should be limited to 3 cm. Meanwhile, ChatGPT-3.5 incorrectly confirmed that frozen section analysis can be used for malignancy assessment and subtyping of soft tissue tumors in question 8. In question 10, ChatGPT-3.5 suggested performing a molecular test to assess the risk of recurrence after sarcoma resection, which actually is not a standard practice according to the German S3 guideline. In the diagnostic section, only ChatGPT-3.5's answer with better accuracy and adequacy than ChatGPT-4's answer was question 12, where ChatGPT-4 answered that the TNM classification can predict the risk of recurrence in GIST patients. For question 6, both chatbots responded with similar answers that a 14G coaxial needle is typically used to perform an image-guided core needle biopsy for soft tissue tumors, although the German S3 guideline suggests that the size of the coaxial needle for biopsy should be at least 16G.

Analysis of treatment questions

For the treatment questions, ChatGPT-3.5 was also defeated by ChatGPT-4 in terms of accuracy and adequacy. In question 22, ChatGPT-3.5 misleadingly suggested that radical lymphadenectomy is not routinely recommended for clear cell sarcomas, rhabdomyosarcomas, epithelioid sarcomas, or myxoid round cell sarcomas with locoregional lymph node involvement. Although ChatGPT-4 responded almost completely correctly to this question, it erroneously equated myxoid round cell sarcoma with myxoid liposarcoma. Both chatbots gave incorrect answers to two key questions (23 and 25) about retroperitoneal sarcomas. The chatbots suggested that adjacent organs without clear evidence of histological infiltration should not be removed during retroperitoneal sarcoma surgery. Meanwhile, they reported that re-resection may be considered for patients who undergo R1 resection for retroperitoneal sarcoma. For the questions on radiotherapy of extremity/trunk sarcomas, ChatGPT-3.5 incorrectly suggested that postoperative radiotherapy can be considered as an alternative to R0 resection for patients who have undergone unplanned R1/R2 resection, to which ChatGPT-4 provided the rational answer (question 33). Both chatbots inappropriately claimed that perioperative radiotherapy has been shown to improve local control but has not shown a significant impact on overall survival in patients who have undergone resection of STSs of the extremities or trunk (question 36). Notably, ChatGPT-3.5 responded with a wrong answer that

Table 2. Assessing accuracy and adequacy responses generated by ChatGPT-3.5 and ChatGPT-4.0

Scores	Accuracy			Adequacy		
(Median (IQR))	ChatGPT-3.5	ChatGPT-4.0	*p* Value	ChatGPT-3.5	ChatGPT-4.0	*p* Value
Overall	5.0 (2.6–6.0)	5.5 (4.1–6)	<0.001	4 (2.5–5.0)	5.0 (4.0–5.0)	<0.001
Diagnostic	4.0 (1.0–5.0)	5.5 (4.0–6.0)	0.007	2.5 (1.8–4.2)	4.5 (4–5)	0.002
Treatment	5.0 (3.0–6.0)	6.0 (4.5–6.0)	<0.001	4.0 (2.5–5.0)	5.0 (4.0–5.0)	<0.001
General	4.8 (2.0–6.0)	5.5 (4.0–6.0)	0.007	3.75 (1.4–4.5)	4.5 (3.0–5.0)	0.003
Extremity/trunk	5.0 (3.0–6.0)	5.8 (4.6–6.0)	0.007	4.25 (2.6–4.9)	5.0 (4.5–5.0	0.003
Retroperitoneal/visceral	4.3(1.1–5.5)	5.3 (1.1–6)	0.102	3.75 (1.1–4.9)	4.5 (1.3–5)	0.102
GIST	6.0 (3.5–6.0)	6.0 (5.0–6.0)	0.581	5.0 (3.0–5.0)	5.0 (4.0–5.0)	0.416
Surveillance	5.0 (2.5–6.0)	5.5 (4.1–6)	<0.001	5.0(4.6–5.0)	5.0 (4.6–5.0)	1.000

IQR, interquartile range; P-values marked with bold indicate statistically significant *p*-values.

pazopanib can be considered a second-line option for the treatment of liposarcoma refractory to previous chemotherapy, to which ChatGPT-4 gave the right suggestion (question 55).

Performance in GIST management

The two chatbots showed similar satisfactory accuracy and adequacy (median accuracy: 6.0 vs. 6.0, $p = 0.581$; median adequacy: 5.0 vs. 5.0, $p = 0.416$). However, ChatGPT-3.5 gave incorrect answers to two questions about tyrosine kinase inhibitor (TKIs) therapy for GIST which ChatGPT-4 answered correctly. For Question 76, ChatGPT-3.5, answered that imatinib was the most effective agent for patients with metastatic or unresectable GIST harboring a D842V mutation in the PDGFRA gene. However, ChatGPT-4 correctly responded that avapritinib was the most effective under such circumstances. In addition, ChatGPT-3.5 suggested that for patients with GIST who have failed multiple kinase inhibitors, clinical trials or compassionate use of newer agents such as ripretinib or avapritinib may be options. However, the German S3 guidelines recommend ripretinib as the first-line treatment.

DISCUSSION

To the best of our knowledge, this study is the first one to compare ChatGPT-3.5 and ChatGPT-4's alignments with the German S3 guideline for adult STS. The results showed that both versions of ChatGPT scored well on answering sarcoma-related questions, but they also provided misleading and potentially life-threatening answers to relevant questions. Additionally, the performance of ChatGPT-4 was superior to that of ChatGPT-3.5 in terms of overall accuracy and adequacy, which is in line with results in previous clinical use cases reported by other authors.[14,22] For instance, ChatGPT-4 achieved a score within the top 10% of participants on a simulated bar exam, whereas ChatGPT-3.5 scored in the bottom 10%.[23] The training cutoff date may significantly affect ChatGPT's accuracy and adequacy on some specific questions, e.g., questions about some newly approved agents, such as avapritinib and ripretinib.

While ChatGPT-4 generally outperformed ChatGPT-3.5, both models gave similar responses to certain questions, especially those related to retroperitoneal/visceral sarcoma, GIST-specific treatment, and surveillance. It could be due to the nature of these questions, which might be based on universally accepted information. Furthermore, potential limitations of AI models could contribute to the lack of variability: both ChatGPT-3.5 and ChatGPT-4 have similar architectures and are not fine-tuned with the latest guidelines. Thus, the models might generate responses which do not reflect the newest standard.

AI is changing biomedical research and healthcare, particularly in the fields of cancer research and treatment. Following the remarkable achievements of AI in laboratory settings, the challenge now is to determine how and when AI can be effectively deployed in the context of everyday clinical practice for cancer patients.[24] Previous studies have shown that ChatGPT can generate accurate, comprehensive, and concise responses to questions from patients with several types of cancers.[20,25,26] However, Valentini et al. observed that responses generated by ChatGPT regarding sarcomas exhibited considerable inconsistency in quality.[18] It is probably because sarcomas are rare diseases and the amount of training data are limited. Although LLM-based chatbots can emulate human language and provide detailed and coherent responses in a timely manner, they can give false and misleading information.[27] It was probable that the chatbot would combine erroneous recommendations with those that were valid, and this was a mistake that even experts were unable to identify easily.[28]

Given its rarity and heterogeneity, sarcomas exemplify many of the difficulties encountered in rare cancers. Centralization is an effective method for enhancing compliance with clinical guidelines and improving patient survival.[4] Nevertheless, it is common for sarcoma patients to lack access to the appropriate information regarding their diagnosis, specialized healthcare facilities, appropriate treatment protocols, and ongoing clinical trials. There is a notable absence of professional expertise among general practitioners in diagnosing and treating sarcomas, which may potentially lead to delays in diagnosis and errors in treatment, which in turn may have adverse effects on the outcomes of patients.[29] Many researchers believe that LLMs, such as ChatGPT, will facilitate access to healthcare services and improve care experiences for professionals and patients. Assuming that ChatGPT can achieve the same level of accuracy as a human expert in responding to patients' questions, it can provide basic medical advice and address operational challenges in low- and middle-income countries and remote areas that suffer from a critical shortage of health professionals.[30] Integrating AI tools like ChatGPT into clinical workflows can significantly support human expertise, especially in the era of personnel shortage. They could take over administrative tasks such as scheduling and patient triage,[31] provide quick access to evidence-based suggestions by analyzing massive amount of data,[32] which enable clinicians to concentrate more on direct patient care. To optimize use of AI tools, AI-generated information needs to be carefully verified and education on their power and limitations are necessary. At the current stage, it must be noted that ChatGPT's performance is not yet optimal, and it is unsafe and unethical to utilize its responses as the basis for directing actual practices, particularly in the absence of supervision by sarcoma experts and a multidisciplinary team.

Considering the variable accuracy of ChatGPT versions in clinical settings, future research should focus on enhancing LLM performance in medical domain. The application of Retrieval-Augmented Generation (RAG) techniques can be explored to improve the performance and understandability of models like ChatGPT. Also, fine-tuning ChatGPT with clinical data can enhance its ability to provide accurate and specialized information in the medical field.

Limitations of the study

Our study has several limitations. First, due to the rapid development in AI technologies, the models available during our research phase may be outdated by the time this report is published. Our data point was based on ChatGPT-3.5 and ChatGPT-4 as on May 5, 2024 when they were the latest models at that time. ChatGPT-4o, released later in mid-May, and subsequent models such as OpenAI-o1, emerged after our study

had already begun or was already under review, highlighting the inherent challenge of maintaining up-to-date assessments in a fast-developing field. Second, different countries have different healthcare systems, clinical practices, and medications. Although the major parts of sarcoma guidelines are largely consistent across countries or regions, there are still some minor variations.[33,34] As a result, there are no one-size-fits-all sarcoma guidelines that can be globally applied, even though it was the first to conduct a systematic literature search and evidence assessment, performed by a scientific research institute.[35] Third, it is uncertain whether ChatGPT is equally effective in answering questions in German, as the LLM was 93% trained with English language texts.[36] It may be beneficial for future studies to consider ChatGPT in multiple languages to evaluate its performance across different linguistic boundaries. Fourth, given the considerable diversity in the presentation, prognosis, and treatment of STS, it is not expected that our sample of 80 questions will be exhaustive in their coverage of these rare diseases. Meanwhile, these questions may not fully encapsulate the intricacies of genuine clinical decision-making in the real world.

Conclusion

This study demonstrates the capability of ChatGPT to align with the German evidence-based S3 guidelines for adult STS. Our findings underscore ChatGPT-4's enhanced ability to process and respond to complex clinical queries with greater accuracy and adequacy than its predecessor, ChatGPT-3.5. Although ChatGPT can provide valuable insights and assist decision-making in sarcoma care, it is not infallible and requires careful supervision by human experts. The integration of ChatGPT in clinical settings should be approached with a balanced understanding of its capabilities and limitations, ensuring that may augment but does not replace human expertise. It is vital that LLMs such as ChatGPT are continually updated and trained to keep pace with rapid advances in medical science and nuanced clinical guidelines. Although ChatGPT shows potential as a supportive tool in sarcoma care in the future, its use must be carefully managed to take advantage of its benefits while minimizing the risks associated with misinformation and overreliance on automated systems.

RESOURCES AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Cui Yang (Cui.Yang@umm.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data: All data reported in this paper will be shared by the lead contact upon request.
- Code: This paper did not report any original code.
- All other items: Additional information is available upon request.

ACKNOWLEDGMENTS

This publication was supported through state funds approved by the State Parliament of Baden-Württemberg for the Innovation Campus Health + Life Science alliance Heidelberg Mannheim. For the publication fee we acknowledge financial support by Heidelberg University. Comments on the manuscript by Dr. Markus Follman, MPH, MSc, Section Director Clinical Practice Guidelines at the German Cancer Society are gratefully acknowledged. The authors used ChatGPT-4.0 to assist with improving readability and performing a grammar check of the manuscript.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study. Material preparation and data collection were conducted by C.L. and C.Y. Statistical analysis was performed by C.L. The initial manuscript draft was authored by C.L., with all authors providing input and revisions to versions. All authors granted final approval of the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no conflict of interest.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Study participants and data collection
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111493.

Received: July 23, 2024
Revised: October 2, 2024
Accepted: November 26, 2024
Published: November 28, 2024

REFERENCES

1. Gamboa, A.C., Gronchi, A., and Cardona, K. (2020). Soft-tissue sarcoma in adults: An update on the current state of histiotype-specific management in an era of personalized medicine. CA A Cancer J. Clin. *70*, 200–229. https://doi.org/10.3322/caac.21605.
2. Gronchi, A., Miah, A.B., Dei Tos, A.P., Abecassis, N., Bajpai, J., Bauer, S., Biagini, R., Bielack, S., Blay, J.Y., Bolle, S., et al. (2021). Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. *32*, 1348–1365. https://doi.org/10.1016/j.annonc.2021.07.006.
3. Blay, J.-Y., Honoré, C., Stoeckle, E., Meeus, P., Jafari, M., Gouin, F., Anract, P., Ferron, G., Rochwerger, A., Ropars, M., et al. (2019). Surgery in reference centers improves survival of sarcoma patients: a nationwide study. Ann. Oncol. *30*, 1143–1153. https://doi.org/10.1093/annonc/mdz124.
4. Blay, J.Y., Penel, N., Valentin, T., Anract, P., Duffaud, F., Dufresne, A., Verret, B., Cordoba, A., Italiano, A., Brahmi, M., et al. (2024). Improved nationwide survival of sarcoma patients with a network of reference centers. Ann. Oncol. *35*, 351–363. https://doi.org/10.1016/j.annonc.2024.01.001.
5. Diessner, B.J., Weigel, B.J., Murugan, P., Zhang, L., Poynter, J.N., and Spector, L.G. (2020). Associations of Socioeconomic Status, Public vs

Private Insurance, and Race/Ethnicity With Metastatic Sarcoma at Diagnosis. JAMA Netw. Open *3*, e2011087. https://doi.org/10.1001/jamanetworkopen.2020.11087.

6. Weaver, R., O'Connor, M., Carey Smith, R., and Halkett, G.K. (2020). The complexity of diagnosing sarcoma in a timely manner: perspectives of health professionals, patients, and carers in Australia. BMC Health Serv. Res. *20*, 711. https://doi.org/10.1186/s12913-020-05532-8.
7. Nijhuis, P.H.A., Schaapveld, M., Otter, R., and Hoekstra, H.J. (2001). Soft tissue sarcoma-Compliance with guidelines. Cancer *91*, 2186–2195. https://doi.org/10.1002/1097-0142(20010601)91:11<2186::AID-CNCR1248>3.0.CO.
8. Sharpless, N.E., and Kerlavage, A.R. (2021). The potential of AI in cancer care and research. Biochim. Biophys. Acta Rev. Canc *1876*, 188573. https://doi.org/10.1016/j.bbcan.2021.188573.
9. Sabeghi, P., Kinkar, K.K., Castaneda, G.D.R., Eibschutz, L.S., Fields, B.K.K., Varghese, B.A., Patel, D.B., and Gholamrezanezhad, A. (2024). Artificial intelligence and machine learning applications for the imaging of bone and soft tissue tumors. Front. Radiol. *4*, 1332535. https://doi.org/10.3389/fradi.2024.1332535.
10. Cao, P., Dun, Y., Xiang, X., Wang, D., Cheng, W., Yan, L., and Li, H. (2024). Machine learning-based individualized survival prediction model for prognosis in osteosarcoma: Data from the SEER database. Medicine *103*, e39582. https://doi.org/10.1097/MD.0000000000039582.
11. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare *11*, 887. https://doi.org/10.3390/healthcare11060887.
12. Yamamoto, A., Koda, M., Ogawa, H., Miyoshi, T., Maeda, Y., Otsuka, F., and Ino, H. (2024). Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial. JMIR Med. Educ. *10*, e58753. https://doi.org/10.2196/58753.
13. Lotto, C., Sheppard, S.C., Anschuetz, W., Stricker, D., Molinari, G., Huwendiek, S., and Anschuetz, L. (2024). ChatGPT Generated Otorhinolaryngology Multiple-Choice Questions: Quality, Psychometric Properties, and Suitability for Assessments. OTO Open *8*, e70018. https://doi.org/10.1002/oto2.70018.
14. Deng, L., Wang, T., Xu, J., Zhai, Z., Zhai, Z., Tao, W., Li, J., Zhao, Y., and Luo, S. (2024). Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int. J. Surg. *110*, 1941–1950. https://doi.org/10.1097/JS9.0000000000001066.
15. Bresler, T.E., Pandya, S., Meyer, R., Htway, Z., and Fujita, M. (2024). From Bytes to Best Practices: Tracing ChatGPT-3.5's Evolution and Alignment With the National Comprehensive Cancer Network® Guidelines in Pancreatic Adenocarcinoma Management. The American Surgeon™ *90*, 2543–2547. https://doi.org/10.1177/00031348241248801.
16. Marchi, F., Bellini, E., Iandelli, A., Sampieri, C., and Peretti, G. (2024). Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses. Eur. Arch. Oto-Rhino-Laryngol. *281*, 2123–2136. https://doi.org/10.1007/s00405-024-08525-z.
17. Matsuoka, M., Onodera, T., Fukuda, R., Iwasaki, K., Hamasaki, M., Ebata, T., Hosokawa, Y., Kondo, E., and Iwasaki, N. (2024). Evaluating the Alignment of Artificial Intelligence-Generated Recommendations With Clinical Guidelines Focused on Soft Tissue Tumors. J. Surg. Oncol. https://doi.org/10.1002/jso.27874.
18. Valentini, M., Szkandera, J., Smolle, M.A., Scheipl, S., Leithner, A., and Andreou, D. (2024). Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? Front. Public Health *12*, 1303319. https://doi.org/10.3389/fpubh.2024.1303319.
19. Hoch, C.C., Wollenberg, B., Lüers, J.-C., Knoedler, S., Knoedler, L., Frank, K., Cotofana, S., and Alfertshofer, M. (2023). ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur. Arch. Oto-Rhino-Laryngol. *280*, 4271–4278. https://doi.org/10.1007/s00405-023-08051-4.
20. Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., and Smith, D.M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern. Med. *183*, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.
21. German Guideline Program in Oncology (German Cancer Society, German Cancer Aid, AWMF): Soft Tissue Sarcoma Long version 1.1, 2022, AWMF Registration Number: 032/044OL, https://www.leitlinienprogramm-onkologie.de/leitlinien/adulteweichgewebesarkome/.; Accessed
22. Lim, Z.W., Pushpanathan, K., Yew, S.M.E., Lai, Y., Sun, C.-H., Lam, J.S.H., Chen, D.Z., Goh, J.H.L., Tan, M.C.J., Sheng, B., et al. (2023). Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine *95*, 104770. https://doi.org/10.1016/j.ebiom.2023.104770.
23. OpenAI; Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., and Altman, S. (2024). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.1371/journal.pdig.0000417.
24. Bhinder, B., Gilvary, C., Madhukar, N.S., and Elemento, O. (2021). Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discov. *11*, 900–915. https://doi.org/10.1158/2159-8290.CD-21-0090.
25. Yalamanchili, A., Sengupta, B., Song, J., Lim, S., Thomas, T.O., Mittal, B.B., Abazeed, M.E., and Teo, P.T. (2024). Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. JAMA Netw. Open *7*, e244630. https://doi.org/10.1001/jamanetworkopen.2024.4630.
26. Yeo, Y.H., Samaan, J.S., Ng, W.H., Ting, P.-S., Trivedi, H., Vipani, A., Ayoub, W., Yang, J.D., Liran, O., Spiegel, B., and Kuo, A. (2023). Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin. Mol. Hepatol. *29*, 721–732. https://doi.org/10.3350/cmh.2023.0089.
27. Van Dis, E.A.M., Bollen, J., Zuidema, W., Van Rooij, R., and Bockting, C.L. (2023). ChatGPT: five priorities for research. Nature *614*, 224–226. https://doi.org/10.1038/d41586-023-00288-7.
28. Chen, S., Kann, B.H., Foote, M.B., Aerts, H.J.W.L., Savova, G.K., Mak, R.H., and Bitterman, D.S. (2023). Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol. *9*, 1459–1462. https://doi.org/10.1001/jamaoncol.2023.2954.
29. Kasper, B., Lecointe-Artzner, E., Wait, S., Boldon, S., Wilson, R., Gronchi, A., Valverde, C., Eriksson, M., Dumont, S., Drove, N., et al. (2018). Working to improve the management of sarcoma patients across Europe: a policy checklist. BMC Cancer *18*, 424. https://doi.org/10.1186/s12885-018-4320-y.
30. Ong, J.C.L., Seng, B.J.J., Law, J.Z.F., Low, L.L., Kwa, A.L.H., Giacomini, K.M., and Ting, D.S.W. (2024). Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. Cell Rep. Med. *5*, 101356. https://doi.org/10.1016/j.xcrm.2023.101356.
31. Levra, A.G., Gatti, M., Mene, R., Shiffer, D., Costantino, G., Solbiati, M., Furlan, R., and Dipaola, F. (2024). A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. Eur. J. Intern. Med. https://doi.org/10.1016/j.ejim.2024.09.017.
32. Yan, Y., Hou, Y., Xiao, Y., Zhang, R., and Wang, Q. (2024). KNOWNET: Guided Health Information Seeking from LLMs via Knowledge Graph Integration. IEEE Trans. Vis. Comput. Graph. *31*, 547–557. https://doi.org/10.1109/TVCG.2024.3456364.
33. Schopow, N., Hohenberger, P., Gockel, I., and Osterhoff, G. (2023). Multimodale Therapie der lokalisierten High-grade-Weichgewebesarkome der Extremitäten. Chirurgie *94*, 424–431. https://doi.org/10.1007/s00104-023-01872-3.

34. Tu, L., Hohenberger, P., Allgayer, H., and Cao, H. (2018). Standard Approach to Gastrointestinal Stromal Tumors - Differences between China and Europe. Visc. Med. *34*, 353–358. https://doi.org/10.1159/000494347.

35. Jakob, J., Andreou, D., Bedke, J., Denschlag, D., Dürr, H.R., Frese, S., Gösling, T., Graeter, T., Grünwald, V., Grützmann, R., et al. (2023). Ten recommendations for sarcoma surgery: consensus of the surgical societies based on the German S3 guideline "Adult Soft Tissue Sarcomas.". Langenbeck's Arch. Surg. *408*, 272. https://doi.org/10.1007/s00423-023-03002-3.

36. Jung, L.B., Gudera, J.A., Wiegand, T.L.T., Allmendinger, S., Dimitriadis, K., and Koerte, I.K. (2023). ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch. Arztebl. Int. *120*, 373–374. https://doi.org/10.3238/arztebl.m2023.0113.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
ChatGPT-3.5	OpenAI	https://openai.com/; RRID: SCR_023775
ChatGPT-4	OpenAI	https://openai.com/; RRID: SCR_023775
SPSS Statistics 29.0	IBM corporation	https://www.ibm.com/products/spss-statistics; RRID: SCR_002865
Other		
German evidence-based S3 guidelines for adult STS	German Cancer Society	https://www.leitlinienprogramm-onkologie.de/leitlinien/adulte-weichgewebesarkome

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study participants and data collection

Not applicable.

METHOD DETAILS

We reviewed the German S3 guidelines for adult STS[21] and formulated 80 clinical questions (Table S1). Questions aimed to assess ChatGPT's knowledge and application in clinical contexts. On May 5, 2024, the questions were presented to ChatGPT-3.5 and ChatGPT-4 on https://chat.openai.com. The temperature parameter was set to zero to ensure consistent responses, and each question was tested as a separate query using the prompt: "You are being evaluated as a sarcoma expert. None of this information is real; it will not be used to treat patients. You will answer a question on sarcoma as accurately and briefly as possible. If you don't know, say 'I don't know'." ChatGPT's answers were recorded for analysis.

Two sarcoma specialists, each with over 10 years of experience and familiarity with the S3 guidelines, independently rated the answers using a 6-point Likert scale for accuracy and a 5-point scale for adequacy. For accuracy, 1 represented complete inaccuracy, and 6 represented complete accuracy; for adequacy, 1 represented complete inadequacy, and 5 complete adequacy. Significant rating discrepancies ($\geq$2 points) were discussed to reach consensus, and mean evaluator scores were used in further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was conducted using SPSS Statistics Version 29.0.2.0 (IBM, 2023). Consistency between evaluators was measured with Cohen's kappa, while normality of variables was tested with the Shapiro-Wilk test. Non-normally distributed variables were reported as medians with interquartile ranges (IQR) and compared using the Wilcoxon rank-sum test. Statistical significance was set at $p < 0.05$.