acta orthopaedica belgica, 2026, 92, online-only

REVIEW

doi.org/10.52628/92.1.14560

Agreement Between AI Language Models and BOOM Chondrosarcoma Consensus Statements: A Comparative Study

A. DÜNKİ[1], Ö. POLAT[1]

[1]*Ortopaedics and Traumatology Department, Umraniye Training and Research Hospital, Istanbul, Turkey.*

Correspondence at: Alper DÜNKİ, Elmalıkent District Adem Yavuz Street No:1 Umraniye Training and Research Hospital, Orthopaedics and Traumatology Department 34764 Umraniye/Istanbul, Turkey. Phone: +90 534 555 06 04 - E-mail: alperdunki@gmail.com

ABSTRACT

Artificial intelligence (AI) is an exciting development makes life easier and solves many problems in daily life. The Birmingham Orthopaedic Oncology Meeting (BOOM) met in January 2024 with 309 participants from 53 countries to discuss the optimal consensus for chondrosarcomas on 21 questions. The aim of this study was to investigate how reliable the expert statements from the BOOM were compared to ChatGPT-4 and DeepseekR1.

21 questions and consensus statements in the section on chondrosarcomas in the BOOM were extracted. The answers were classified according to the level of evidence and consensus status, taking into account the consensus strength category determined in the meeting. Each statement were written separately for the ChatGPT-4 and DeepseekR1. Consensus questions and answers were written for the AI modules and they were asked to interpret these expressions as ‘’strongly disagree, disagree, undecided, agree or strongly agree’’.

BOOM participants reached a strong consensus on 19 questions. The number of people who accepted the proposition for 1 question was 52% and no consensus was reached. ChatGPT-4 and DeepseekR1 responded ‘’disagree’’ for a same question. The level of evidence for that question was ‘’low to moderate’’ and a strong consensus was reached. A significant relationship was found between the responses of ChatGPT-4 and DeepseekR1.

ChatGPT-4 and Deepseek expressed more positive opinions in the answers with high levels of evidence, while BOOM participants were able to make stronger consensus decisions by combining their clinical observations with literature knowledge, regardless of level of evidence.

Keywords: ChatGPT, Deepseek, artificial intelligence, orthopaedic oncology, chondrosarcoma.

INTRODUCTION

Artificial intelligence (AI) is an exciting development that makes life easier and solves many problems in daily life. In America, 80% of patients use websites and AI modules to access information about health[1]. AI, which is frequently used in modern medicine, contributes to a positive development in clinical decision-making and patient care quality. It is also frequently used for fast and fluent evaluation in literature reviews[2,3]. ChatGPT-4 is frequently used in the literature in disease diagnosis and treatment. The AI model called ChatGPT-4 entered our lives with its first version in November 2022[4]. ChatGPT-4, which has a more detailed proposal in academic studies with current updates, is used. DeepseekR1 is a newly defined AI model that was first introduced in January 2025, and a module that dominates the literature was developed with the DeepseekR1 version with the deepthink mode[5]. In this way, the AI modules that can be scanned in literature have been enriched. The Birmingham Orthopaedic Oncology Meeting (BOOM) met in January 2024 with 309 participants from 53 countries to discuss the optimal consensus for chondrosarcomas on 21 questions. Participants had been caring for orthopaedic tumours for an average of 15 years (range 1-40 years) and saw an average of 97 bone sarcomas per year. The average number of patients seen by participants per year multiplied by the number of participants was estimated to give a combined experience of treating 30,000 bone sarcomas per year. This estimate represents approximately 450 new bone

sarcomas seen in the UK each year and is estimated to be equivalent to the global experience of the meeting, which is equivalent to 66 years of UK experience[6]. The aim of this study was to investigate how reliable the expert experience statements from the BOOM consensus meeting were compared to ChatGPT-4 and DeepseekR1.

METHODS

21 questions and consensus answers in the section on chondrosarcomas in the BOOM consensus meeting were extracted. The answers were classified according to the level of evidence and consensus status, taking into account the consensus strength category determined in the meeting (Table I).

The levels of evidence in the BOOM consensus meeting were divided into 5 groups from highest to lowest as ‘’high’’, ‘’moderate to high’’, ‘’moderate’’, ‘’low to moderate’’ and ‘’low’’. The consensus strength was determined in 5 groups according to the positive opinions given by the participants in the meeting to the answers. A 100% positive opinion was defined as ‘’unanimous’’, a positive opinion between 80-99% was defined as ‘’strong’’, between 70-79% as ‘’moderate’’, between 60-69% as ‘’weak’’ and below 59% as ‘’no consensus’’.

In the meeting; ‘’Strong consensus’’ was achieved for 19 out of 21 questions, moderate consensus was achieved for one question, and consensus was not achieved for one question (Table II). Each consensus question and answer were typed into the ChatGPT-4 and DeepseekR1 AI modules. Respondents were asked to rate their responses on a Likert scale from highest to lowest: “strongly agree, agree, undecided, disagree, or strongly disagree”. After each statement, the AI module was closed and restarted, and the conversation history was deleted. In case of inconsistent answers, the questions were asked again with the same expression. The AI modules were asked to briefly provide their answers and to definitely mention these expressions because ChatGPT-4 and DeepseekR1 had very detailed answers. Data were collected in February 2025. Statistical analysis was performed with the IBM SPSS v23 program. Compatibility of the data with normal distribution was assessed using the Shapiro Wilk test. The relationship between variables with non-normal distribution was assessed using the Spearman correlation test. A p value of <0.05 was considered significant.

Ethical approval was not required because this study analysed published consensus statements and AI-generated text only, without involving human subjects or identifiable patient data.

RESULTS

The relationship between the consensus questions and answers answered by 309 participants in BOOM, which were listed as ‘’low, low to moderate, moderate, moderate to high and high’’ according to the level of evidence, and the consensus strength category that emerged in the meeting was examined. In 7 out of 21 questions, the level of evidence was shown as ‘’moderate’’, in 2 of them the level of evidence was shown as ‘’low to moderate’’ and in 12 of them the level of evidence was shown as ‘’low’’. When the consensus strength data were examined, while ‘’unanimous’’ and ‘’weak’’ consensus was not seen at all, in 19 of them ‘’strong’’, in 1 of them ‘’moderate’’, and in 1 of them ‘’no consensus’’ was seen.

While the only question where no consensus was formed was seen as the level of evidence as ‘’moderate’’, strong consensus was seen in 13 questions with ‘’low’’ and ‘’low to moderate’’ levels of evidence. A ‘’moderate’’ consensus was also seen for a question with a ‘’low’’ level of evidence. Statistically, no significant correlation was found between the BOOM consensus strength and the level of evidence (p:0.690).

The level of evidence was compared separately for ChatGPT-4 and DeepseekR1 responses. ChatGPT-4 responded to 5 questions with ‘’strongly agree’’, 12 questions with ‘’agree’’, 3 questions with ‘’undecided’’ and 1 question with ‘’disagree’’, while it did not respond to any question with ‘’strongly

Table I. — BOOM Consensus Strength Categories.

BOOM Consensus Strength Categories	
Unanimous (%100)	Unanimous Consensus
Super Majority (%80-99)	Strong Consensus
Large Majority (%70-79)	Moderate Consensus
Majority (%60-69)	Weak Consensus
Simple Majorty (%50.1-59)	No Consensus

Table I. — BOOM Consensus Questions, Consensus Levels, ChatGPT and Deepseek Responses.

Questions	Evidence Level	n	Results (%)			Consensus Level	ChatGPT	Deepseek
			Agree	Disagree	Abstain			
Radiology of cartilage tumours								
Which imaging feature gives the best positive/negative predictive value for differentiating an enchondroma from an atypical chondroid tumour (ACT)/ chondrosarcoma?	Moderate	225	95	2	3	Strong Consensus (%95)	Agree	Strongly Agree
Can chondrosarcoma be safely diagnosed by radiology alone using radiology classifications e.g. BACTIP (Birmingham Atypical Cartilaginous Tumour Imaging Protocol)?	Moderate	235	82	14	2	Strong Consensus (%85)	Agree	Agree
Surveillance of chondrosarcoma								
What is the optimal clinical and radiological surveillance following chondrosarcoma resection? Should we stratify by risk?	Low	224	87	9	3	Strong Consensus (%91)	Agree	Agree
Is it safe to undertake radiological surveillance in ACT? What is optimal interval between scans and when should we intervene?	Low / Moderate	242	89	8	3	Strong Consensus (%92)	Disagree	Disagree
Intraosseous ACT/chondrosarcoma								
Do purely intraosseous central cartilage tumours/ACT/ chondrosarcoma metastasize?	Low	218	92	5	3	Strong Consensus (%95)	Agree	Agree
How should we treat intraosseous ACT/ chondrosarcoma?	Moderate	230	49	45	6	No Consensus (%52)	Agree	Agree
Is it safe to avoid biopsy in radiologically typical chondrosarcomas/ACT?	Low / Moderate	233	82	15	3	Strong Consensus (%85)	Strongly Agree	Agree
Locally recurrent disease								
Does local recurrence influence the prognosis for chondrosarcoma?	Moderate	229	97	1	1	Strong Consensus (%99)	Strongly Agree	Agree
How aggressive should we be in treating locally recurrent disease in chondrosarcoma?	Moderate	215	95	3	2	Strong Consensus (%97)	Agree	Agree
Dedifferentiated chondrosarcoma								
How aggressive should we be with surgery on dedifferentiated chondrosarcoma?	Low	215	97	1	2	Strong Consensus (%99)	Agree	Agree
Should we routinely use adjuvant/ neoadjuvant chemotherapy with localized de- differentiated chondrosarcoma?	Low	230	87	7	6	Strong Consensus (%93)	Undecided	Agree
Surgical margins								
What is a wide margin in chondrosarcoma?	Low	213	74	22	4	Moderate Consensus (%77)	Agree	Strongly Agree
Should we vary the attempted surgical margin depending on grade of chondrosarcoma?	Low	211	99	1	0	Strong Consensus (%99)	Agree	Agree

Table I. — BOOM Consensus Questions, Consensus Levels, ChatGPT and Deepseek Responses - continued.

Questions	Evidence Level	n	Results (%)			Consensus Level	ChatGPT	Deepseek
			Agree	Disagree	Abstain			
Treatment of inadvertent margins								
Do intralesional margins for high- grade chondrosarcoma increase risk of poor oncological outcomes?	Moderate	215	97	1	2	Strong Consensus (%99)	Strongly Agree	Strongly Agree
What is the optimal treatment following an inadvertent intralesional margin of a high- grade chondrosarcoma?	Low	216	92	7	1	Strong Consensus (%93)	Undecided	Agree
Pathological fractures								
Does pathological fracture influence the outcome for chondrosarcoma?	Low	220	97	1	2	Strong Consensus (%99)	Agree	Agree
Is limb salvage safe in patients presenting with a pathological fracture through chondrosarcoma?	Low	214	91	6	3	Strong Consensus (%94)	Agree	Agree
Pelvic chondrosarcomas								
Do pelvic chondrosarcomas behave more aggressively and therefore should they be treated more aggressively?	Moderate	209	92	4	4	Strong Consensus (%94)	Strongly Agree	Strongly Agree
Does navigated surgical resection (with jigs or computer navigation) of chondrosarcoma of pelvis result in better oncological outcomes?	Low	211	93	4	3	Strong Consensus (%96)	Undecided	Agree
Adjuvant treatment							Strongly Agree	Agree
What is the role of adjuvant therapy (radiotherapy/proton beam therapy/ carbon ion/chemotherapy) in conventional chondrosarcoma?	Low	209	86	10	4	Strong Consensus (%90)		
Is there a role for alternate treatments in chondrosarcoma (e.g. cryoablation/ radiofrequency ablation (RFA)/ extracorporeal irradiation and reimplantation (ECRI))?	Low	209	90	6	4	Strong Consensus (%96)	Agree	Agree

disagree'' (Table III). The question that ChatGPT-4 and DeepseekR1 answered ''disagree'' is ''Is it safe to undertake radiological surveillance in ACT?'' with a ''strong consensus'' level in BOOM What is the optimal interval between scans and when should we intervene?''. The 3 questions it responded to with ''undecided'' were also ''low'', which is the lowest level of evidence. 3 of the 5 questions it responded to with ''strongly agree'' had a ''moderate'' level of evidence. No statistically significant correlation was found between the level of evidence and ChatGPT-4 and DeepseekR1 responses (p:0.077 and p:0.210).

When the DeepseekR1 AI module responses were examined; He gave the answers ''strongly agree'' to 4 questions, ''agree'' to 16 questions, ''disagree'' to 1 question, and did not use the answers ''strongly disagree'' or ''undecided'' to any question. 3 out of the 4 questions he gave the answer ''strongly agree'' had a ''moderate'' level of evidence. When statistically examined, $p<0.05$ was seen and a significant relationship was seen between DeepseekR1 answers and the level of evidence of the questions. ChatGPT-4 answers were compared with consensus answers. Of the 19 questions that had a ''strong'' level of consensus, Chat GPT gave the answers ''strongly agree'' to 5, ''agree'' to 10, and the only question he gave the answer ''disagree'' was in this group. ChatGPT-4 gave the answer ''agree'' to the question that did not have a consensus and had a ''moderate'' level of consensus. No statistically significant correlation was found between consensus strength and ChatGPT-4 and DeepseekR1 responses (p:0.897 and p:0.305).

DeepseekR1 responses were compared with consensus responses. DeepseekR1 responded to 3

Table III. — ChatGPT and Deepseek Responses.

Responses	ChatGPT	Deepseek
Strongly Agree	5	4
Agree	12	16
Undiceded	3	0
Disagree	1	1
Strongly Disagree	0	0

Table IV. — Correlation of groups.

Correlation	p-Value	rho
Consensus Strength – Level of Evidence	[a]0.690	-
Consensus Strength – ChatGPT	[a]0.897	-
Consensus Strength – Deepseek	[a]0.305	-
Level of Evidence – ChatGPT	[a]0.077	-
Level of Evidence – Deepseek	[a]0.210	-
ChatGPT – Deepseek	[a]0.036*	0.460
[a]: Spearson's correlation test; *: Statistically significant.		

out of 19 questions that reached a consensus level of "strong" with "strongly agree" and 15 with "agree", while the only question it responded "disagree" was in this group. While the question that did not reach consensus was answered with "agree", DeepseekR1 responded with "strongly agree" to the question that reached a consensus level of "moderate". No statistically significant correlation was found between DeepseekR1 and consensus responses.

Finally, the responses given by ChatGPT-4 and DeepseekR1 AI modules were compared. The only question that both AI modules responded with "disagree" was the same question. DeepseekR1 responded with "agree" to all 3 questions that ChatGPT-4 responded with "undecided". When the relationship between DeepseekR1 and ChatGPT-4 responses was examined, a statistically significant mid-level positive correlation was observed between the response (p:0.036, rho:0.460) (Table IV).

DISCUSSION

In this study, it was seen that the answers of the AI modules were consistent with each other and the rate of agreement with the questions increased as the level of evidence increased. Most of the DeepseekR1 and ChatGPT-4 answers were found to be similar to the consensus levels. The answer given to the question about the safety of radiological screening and the screening interval for atypical chondromatous tumors (ACT) for which no consensus was reached was defined as ''agree'' by both AI modules. When the literature is examined, it is also recommended that curettage grafting or active surveillance be performed during the follow-up period when radiological follow-up is sufficient for ACTs[7,8]. '' Limited evidence suggests that the risk of metastatic disease from ACTs is very low and that radiological surveillance is safe in the medium term for ACTs, but there is no protocol for the duration or interval of follow-up.'' There was no consensus for the answer given[6].

The ''Is it safe to undertake radiological surveillance in ACT?'' with a ''strong consensus'' level in BOOM What is the optimal interval between scans and when should we intervene?'' question, the need for follow-up has been shown in the literature, but there is no clear information about the follow-up period[9]. ChatGPT-4 and DeepseekR1 stated in their responses that the follow-up period was not clear in the literature and wrote that they determined the answer as ''disagree''.

When the responses in BOOM were examined, although 100% consensus could not be achieved, a ''strong'' consensus response corresponding to 80-99% consensus was given in 19 questions. While ChatGPT-4 gave 5 ''strongly agree'' and 12 ''agree'' responses, DeepseekR1 gave 4 ''strongly agree'' responses and 16 ''agree'' responses, which was seen to be more compatible with the consensus responses.

Guardiani et al. asked 20 questions about anterior cruciate ligament reconstruction to Google web and ChatGPT-4, and evaluated the scientific nature and accuracy of the questions and answers with the Rothwell system Flesch-Kincaid grade level. As a

result, ChatGPT-4's answers were found to be much more accurate and more precise[10]. Zhou et al. asked ChatGPT-4 and DeepseekR1 AI modules to create an educational module for lumbar discectomy, spinal fusion, and decompressive laminectomy from 3 different spine surgeries. They compared them with DISCERN score and Flesch-Kincaid Grade Level (FKGL) and saw the best results in ChatGPT-4 and then in DeepseekR1 module[11]. In our study, in accordance with the literature, the level of evidence of BOOM consensus questions were found to be consistent with ChatGPT-4 and DeepseekR1 answers.

Hurley et al. asked ChatGPT-4 questions and expert statements from a consensus meeting on anterior shoulder instabilities and compared the answers with the answers of orthopedic surgeons; they found a limited correlation between them[12]. In our study, we did not find a statistically significant relationship between ChatGPT-4 and consensus data, but the agreement rates of ChatGPT-4 and DeepseekR1 AI modules increased with the level of evidence.

Cuthbert et al. asked ChatGPT-4 the 1st stage questions of the Royal College of Fellowship in England in 2023, and it was observed that ChatGPT-4 gave successful answers to basic science questions, while ChatGPT-4 was inadequate in trauma questions requiring clinical experience[13]. In line with the literature, we also observed that AI modules agreed more on questions with a high level of evidence. Similarly, they were more hesitant in statements with a low level of evidence and for which BOOM participants had a strong consensus based on their clinical observations.

Chester et al. asked ChatGPT-4 to interpret the responses it received about 14 statements with consensus and 9 statements without consensus in a consensus on surgical site infections in pediatric spinal surgeries[14]. ChatGPT-4 stated that it mostly strongly agreed with the statements with consensus. This study demonstrated an acceptable correlation between the statement of participating surgeons and their ChatGPT-4 responses.

Zaboli et al. found a similar result when they examined ChatGPT-4's ability to classify patients; their research showed that ChatGPT-4's ability to provide correct answers decreased as the complexity of the questions increased[15]. AI modules create their own library by scanning all the information on the web. These sources include publicly available data on the web, licensed information from third parties, and information from human trainers. This provides AI modules with a rich repository of information, but the internet is a medium where everyone can express their opinions, increasing the possibility that modules may have accessed incorrect and misleading information.

Current AI models generate answers based on patterns in published texts. They lack tacit clinical experience and cannot weigh nuanced clinical context as expert panels do. As a result, their recommendations often reflect the average of published data rather than actual practice in specialised centres.

The discrepancy in the question on radiological surveillance in ACT highlights this gap. Experts concluded that medium-term surveillance is safe because metastatic risk is very low, even though no standard protocol exists for scan intervals or intervention criteria. In contrast, the AI model acknowledged safety but could not suggest follow-up intervals or when to intervene, merely stating that no protocols exist. This reflects its reliance on published evidence without the tacit clinical judgement experts apply in areas with limited data. Thus, even when AI aligns with general principles, it may fail to provide practical, clinically actionable guidance in nuanced settings.

Importantly, emerging clinical data support the safety of active surveillance for ACT[16]. In a cohort of 128 central cartilaginous tumours in long bones followed for a mean of 50 months, 87% remained stable or regressed on MRI and none progressed to high-grade chondrosarcoma[17]. Earlier, in a smaller series of 49 conservatively managed patients, only ~6% required surgery and the authors recommended annual radiologic follow-up for asymptomatic lesions irrespective of size[18]. These data illustrate why expert consensus favours surveillance rather than aggressive intervention — a nuance that the AI model failed to translate into concrete recommendations.

Overall, while AI may help summarise evidence, it should not replace structured expert consensus — especially in rare tumours such as chondrosarcoma where experience matters[19,20].

This study has limitations. It includes only 21 chondrosarcoma questions, limiting generalisability. AI outputs are time-specific, generated in February 2025, and may change with future updates. Finally, we evaluated agreement with existing consensus statements, not the ability of AI to generate consensus de novo. Our findings therefore reflect alignment with established guidance rather than true consensus-building capacity. Secondly, Static snapshot of AI models in February 2025; LLM behaviour is version- and date-dependent, so reproducibility over time is uncertain[21]. AI models were not tested for their ability

to generate consensus statements de novo, only for their agreement with given statements; this evaluates alignment, not autonomous reasoning. AI training data and access to paywalled sarcoma literature are not fully transparent and may bias responses[22]. AI modules have the ability to receive and process information from all sources in all languages, but it is unclear how many of the AI modules' data partnerships are in which languages and their level of scientific evidence. This raises questions about whether AI modules have access to quality data sources published in all languages. In addition, it is not known to what extent AI modules have access to the literature information that is available for a fee on the internet. This probably limits access to a significant portion of peer-reviewed and paid content.

Finally, BOOM is a consensus meeting where the current literature on the diagnosis, follow-up and treatment of chondrosarcomas in orthopedic oncology is combined with the clinical observations of the participants. A strong consensus was reached on 19 out of 21 questions that were on the minds of patients with chondrosarcoma.

The questions and answers in this meeting were asked to ChatGPT-4 and DeepseekR1, which are AI modules that receive opinions on every subject today. The AI modules were asked for their opinions on the consensus questions with short answers. In the study, as the level of evidence for the answers increased, the degree of agreement of ChatGPT-4 and DeepseekR1 with the answers increased. However, no significant relationship was observed between the answers when compared to the consensus participants. As a result; ChatGPT-4 and DeepseekR1 expressed more positive opinions in the answers with high levels of evidence, while BOOM participants were able to make stronger consensus decisions by combining their clinical observations with literature knowledge, regardless of the level of evidence.

Acknowlegments: No funding received from any institution. No potential conflict of interest relevant to this article was reported.

Credit roles: Alper Dünki: Conceptualization; Data curation; Writing – review & editing ; Ömer Polat: Software; Supervision; Validation; Visualization.

REFERENCES

1. Calixte, R.; Rivera, A.; Oridota, O.; Beauchamp, W.; Camacho-Rivera, M. Social and Demographic Patterns of Health-Related Internet Use Among Adults in the United States: A Secondary Data Analysis of the Health Information National Trends Survey. Int. J. Environ. Res. Public Health 2020, 17, 6856. https://doi.org/10.3390/ijerph17186856
2. Picton B, Andalib S, Spina A, Camp B, Solomon SS, Liang J, Chen PM, Chen JW, Hsu FP, Oh MY. Assessing AI Simplification of Medical Texts: Readability and Content Fidelity. Int J Med Inform. 2025 Mar;195:105743. doi: 10.1016/j.ijmedinf.2024.105743. Epub 2024 Dec 1. PMID: 39667051.
3. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in Orthopaedics and sports medicine. J Exp Orthop. 2023 Jul 26;10(1):74. doi: 10.1186/s40634-023-00642-8. PMID: 37493985; PMCID: PMC10371934.
4. OpenAI. OpenAI website. Available from: https://openai.com. Accessed June 30, 2024.
5. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, Temsah MH. Deepseek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier. Cureus. 2025 Feb 18;17(2):e79221. doi: 10.7759/cureus.79221. PMID: 39974299; PMCID: PMC11836063.
6. Jeys LM, Morris GV, Kurisunkal VJ, Botello E, Boyle RA, Ebeid W, Houdek MT, Puri A, Ruggieri P, Brennan B; BOOM Consensus Meeting Participants; Laitinen MK. Identifying consensus and areas for future research in chondrosarcoma: a report from the Birmingham Orthopaedic Oncology Meeting. Bone Joint J. 2025 Feb 1;107-B(2):246-252. doi: 10.1302/0301-620X.107B2.BJJ-2024-0643.R1.PMID: 39889743.
7. Deckers C, Rooy JWJ, Flucke U, Schreuder HWB, Dierselhuis EF, Geest ICMV. Midterm MRI Follow-Up of Untreated Enchondroma and Atypical Cartilaginous Tumors in the Long Bones. Cancers (Basel). 2021 Aug 13;13(16):4093. doi: 10.3390/cancers13164093. PMID: 34439246; PMCID: PMC8393576.
8. Brown MT, Gikas PD, Bhamra JS, Skinner JA, Aston WJ, Pollock RC, Saifuddin A, Briggs TW. How safe is curettage of low-grade cartilaginous neoplasms diagnosed by imaging with or without pre-operative needle biopsy? Bone Joint J. 2014 Aug;96-B(8):1098-105. doi: 10.1302/0301-620X.96B8.32056. PMID: 25086127.
9. Chung BM, Hong SH, Yoo HJ, Choi JY, Chae HD, Kim DH. Magnetic resonance imaging follow-up of chondroid tumors: regression vs. progression. Skeletal Radiol. 2018 Jun;47(6):755-761. doi: 10.1007/s00256-017-2834-z. Epub 2017 Dec 3. PMID: 29197957.
10. Gaudiani MA, Castle JP, Abbas MJ, Pratt BA, Myles MD, Moutzouros V, Lynch TS. ChatGPT-4 Generates More Accurate and Complete Responses to Common Patient Questions About Anterior Cruciate Ligament Reconstruction Than Google's Search Engine. Arthrosc Sports Med Rehabil. 2024 Apr 9;6(3):100939. doi: 10.1016/j.asmr.2024.100939.PMID: 39006779; PMCID: PMC11240040.
11. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025 Mar 13;198:105871. doi: 10.1016/j.ijmedinf.2025.105871. Epub ahead of print. PMID: 40107040.
12. Hurley ET, Matache BA, Wong I, et al. Anterior shoulder instability. Part I—diagnosis, nonoperative management, and bankart repair—an international consensus statement. Arthroscopy. 2022; 38:214–223.e7.
13. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. 2023 Sep 21;99(1176):1110-1114. doi: 10.1093/postmj/qgad053. PMID: 37410674.
14. Chester AN, Mandler SI. A Comparison of ChatGPT and Expert Consensus Statements on Surgical Site Infection Prevention in

High-Risk Paediatric Spine Surgery. J Pediatr Orthop. 2025 Jan 1;45(1):e72-e75. doi: 10.1097/BPO.0000000000002781. Epub 2024 Aug 30. PMID: 39210518.
15. Zaboli A, Brigo F, Sibilio S, et al. Human intelligence versus ChatGPT: who performs better in correctly classifying patients in triage? Am J Emerg Med. 2024;79:44–47.
16. Quiriny M, Gebhart M. Chondrosarcoma of the spine: a report of three cases and literature review. Acta Orthop Belg. 2008 Dec;74(6):885-90. PMID: 19205342.
17. Woltsche JN, Smolle MA, Szolar D, Leithner A. Follow-up analysis of lesion characteristics of enchondromas and atypical cartilaginous tumours of the knee and shoulder region on MRI. Eur Radiol. 2025;35(5):2935-2945. doi:10.1007/s00330-024-11106-7
18. Deckers C, Schreuder BH, Hannink G, de Rooy JW, van der Geest IC. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumors in the long bones. J Surg Oncol. 2016 Dec;114(8):987-991. doi: 10.1002/jso.24465. Epub 2016 Oct 3.
19. Gunay C, Atalar H, Hapa O, Basarir K, Yildiz Y, Saglik Y. Surgical management of grade I chondrosarcoma of the long bones. Acta Orthop Belg. 2013 Jun;79(3):331-7. PMID: 23926738.
20. Makar GS, Udoeyo IF, Bowen TR. Non-Operative Treatment of patients with Chondrosarcoma: An analysis of patients who refused cancer-directed surgery or patients contraindicated to surgery. Acta Orthop Belg. 2024 Dec;90(4):745-758. doi: 10.52628/90.4.12611. PMID: 39869879.
21. Hu X, Niemann M, Kienzle A, Braun K, Back DA, Gwinner C, Renz N, Stoeckle U, Trampuz A, Meller S. Evaluating ChatGPT responses to frequently asked patient questions regarding periprosthetic joint infection after total hip and knee arthroplasty. Digit Health. 2024 Aug 9;10:20552076241272620. doi: 10.1177/20552076241272620. PMID: 39130521; PMCID: PMC11311159.
22. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al.; FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025 Feb 5;388:e081554. doi: 10.1136/bmj-2024-081554. Erratum in: BMJ. 2025 Feb 17;388:r340. doi: 10.1136/bmj.r340. PMID: 39909534; PMCID: PMC11795397.