

## Research Article

## Evaluation of AI-Based Chatbots in Responding to Frequently Asked Online Questions About Perthes Disease: A Comparison of ChatGPT-5.1, Gemini 3, DeepSeek V3.2

Tahsin Olgun BAYRAKTAR<sup>1</sup> Nazım ERKURT<sup>1</sup> Ali YÜCE<sup>1</sup> Mustafa YERLİ<sup>1</sup> Baran ÖZDEMİR<sup>1</sup> Ahmet Murat ÇÖREKÇİ<sup>1</sup><sup>1</sup> Sağlık Bilimleri Üniversitesi, İstanbul Prof. Dr. Cemil Taşçıoğlu Şehir Sağlık Uygulama ve Araştırma Merkezi, Cerrahi Tıp Bilimleri Bölümü, Ortopedi ve Travmatoloji Anabilim Dalı

## ARTICLE INFO

## Article history:

Submitted June 11

Accepted July 06

Publication September 30

## Keywords:

Legg-Calvé-Perthes disease

chatbot

artificial intelligence

online health information

## ORCID iDs of the Corresponding:

Tahsin Olgun Bayraktar, MD  
0000-0001-6544-1710

## ABSTRACT

**Objective:** Legg-Calvé-Perthes disease is a serious orthopedic condition that can lead to permanent disability. Today, families increasingly turn to artificial intelligence-based chatbots in their online health information search. This study aimed to test the hypothesis that three commonly used chatbots can provide accurate and understandable answers to questions about Perthes disease, and to compare their responses in terms of accuracy, comprehensiveness, readability, and user understandability.**Methods:** On December 10, 2025, the AI-based chatbots GPT-5.1 (OpenAI), Gemini 3 (Google), and DeepSeek V3.2 (DeepSeek) were asked 11 questions from the "Frequently Asked Questions" section of the Perthes Kids Foundation website. A separate conversation was started for each question, using the standard prompt: "I am a parent of a child diagnosed with Perthes disease and I would like to ask...". The answers were evaluated by two independent pediatric orthopedic specialists using AIRES, PEMAT-P, Flesch Reading Ease (FRE), and Flesch-Kincaid Grade Level (FKGL) indices and were statistically compared.**Results:** All chatbots produced excellent-quality answers according to the AIRES evaluation. A moderate and statistically significant agreement was observed between observers for AIRES (ICC=0.501, p<0.001). Gemini provided answers with lower understandability according to PEMAT-P scores (p = 0.005). No significant differences were found between the chatbots in readability analyses; most answers were at a high school-level readability.**Conclusion:** This study shows that current chatbots can provide accurate, comprehensive, up-to-date, and understandable answers to questions about Perthes disease. Their interactive nature may allow them to differ from traditional online health information platforms and to respond more effectively to the multidimensional information needs of patients and families. In the future, integrating chatbots with clinical findings may further enhance personalized information delivery.

## Doi:

<https://doi.org/10.5281/zenodo.22826931>Nazım ERKURT, MD  
0000-0003-4329-1437Ali YÜCE, MD  
0000-0003-3751-6644Mustafa YERLİ, MD  
0000-0002-2708-5812Baran ÖZDEMİR, MD  
0000-0002-8484-7844Ahmet Murat ÇÖREKÇİ, MD  
0009-0008-7505-8449

## Introduction

Legg-Calvé-Perthes disease is a serious pediatric orthopedic condition that may show a progressive course and may lead to permanent disability (Gilbert et al., 2025). Its clinical course, which can last for years, may negatively affect the child's daily activities, school life, and social participation, while also causing uncertainty and future-related anxiety in parents (Leo et al., 2019). The long and unpredictable nature of the disease may markedly increase disease-related information seeking among patients and their families (Kubb & Foran, 2020).

Today, families increasingly access health-related information through online platforms (Kubb & Foran, 2020). However, many studies have suggested that online health information is often incomplete, inaccurate, or unreliable, and that this situation may negatively influence patient and family attitudes (Lock & Baker, 2022; Winship et al., 2014). In recent years, this information-seeking behavior has shifted toward artificial intelligence-based chatbots due to their rapid and easy accessibility (Yun & Bickmore, 2025). Although chatbots are claimed by their developers to have the potential to provide evidence-based information, the accuracy of their responses and their understandability for users remain important areas of uncertainty (Pan et al., 2023; Walker et al., 2023).

In this study, the hypothesis that three commonly used chatbots may provide accurate and understandable answers to questions posed about Perthes disease was tested, and the responses of these systems were compared in terms of accuracy, comprehensiveness, readability, and understandability.

## Methods

In the design and reporting process of this cross-sectional study, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines recommended for observational research were followed (von Elm et al., 2007). On 10 December 2025, using the Google Chrome browser with cache and cookies cleared, all questions from the "frequently asked questions" section of the official Perthes Kids Foundation website ([www.pertheskids.org](http://www.pertheskids.org)) were posed to the artificial intelligence-based chatbots GPT-5.1 (OpenAI), Gemini 3 (Google), and DeepSeek V3.2 (DeepSeek) (Table 1). The questions were selected because they reflect issues frequently raised by families in pediatric orthopedic clinical practice and address not only medical information but also daily life, school adaptation, and social interactions.

## Corresponding author:

Tahsin Olgun Bayraktar, MD  
olgun.bayraktar@yahoo.com**Cite this article:** Bayraktar, T. O., Erkurt, N., Yüce, A., Yerli, M., Özdemir, B., & Çörekci, A. M. (2026).Evaluation of AI-based chatbots in responding to frequently asked online questions about Perthes disease: A comparison of ChatGPT-5.1, Gemini 3, DeepSeek V3.2. *Journal of Baltalimani*, 2(3), 60–64.

This approach was intended to evaluate not only the chatbots' capacity to generate medical responses but also their potential to provide information regarding the multidimensional effects of the disease.

The chatbots included in the study were selected based on their popularity, ease of access, and widespread public use for obtaining health-related information. A separate conversation was initiated for each question; to establish a standard context, the prompt "I am a parent of a child diagnosed with Perthes disease and I would like to ask..." was used, and the questions were added to the continuation of this statement. All questions were directed to each chatbot using the same prompt and only once.

All responses to the total of 11 questions were evaluated independently by two pediatric orthopedic surgeons. During the evaluation process, the Artificial Intelligence Response Evaluation Scale (AIRES), the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), the Flesch Reading Ease (FRE) score, and the corresponding Flesch-Kincaid Grade Level (FKGL) indices were used. The obtained data were compared statistically.

AIRES is a measurement tool developed by two senior pediatric orthopedic surgeons and is intended to quantitatively evaluate the quality of medical informational responses generated by artificial intelligence, consisting of four main domains. These domains include accuracy, comprehensiveness, understandability, and up-to-dateness of information. Each domain was scored using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The evaluations were performed by two senior pediatric orthopedic surgeons; the same evaluators re-scored the same responses one month later, and the mean of the two assessments was accepted as the final score. Total AIRES scores were classified as <8 insufficient, 9–12 low quality, 13–16 good quality, and 17–20 excellent quality responses (Table 2).

The understandability and actionability of the responses for patients and parents were analyzed using PEMAT-P, a standard evaluation tool developed for written patient education materials. In PEMAT-P, each item was rated as "yes" (1 point), "no" (0 points), or "not applicable"; understandability and actionability scores were calculated as the percentage of "yes" responses for the relevant items. Higher scores indicate that the material may be more understandable and actionable. PEMAT-P results were interpreted as insufficient quality below 70% and sufficient quality at 70% or above. As PEMAT-P is a structured tool with established validity, interobserver agreement analysis was not additionally performed in this study.

Text readability was evaluated using the Flesch Reading Ease (FRE) index. FRE generates a score between 0 and 100 based on sentence length and the number of syllables per word, with higher scores indicating that the text may be easier to read. In addition, Flesch-Kincaid Grade Level (FKGL) values corresponding to the FRE scores were calculated to report the approximate educational level required to understand the texts. FRE scores were classified as follows: 90–100 very easy, 80–89 easy, 70–79 fairly easy, 60–69 standard, 50–59 fairly difficult, 30–49 difficult, and 0–29 very difficult readability levels. FKGL scores were interpreted to reflect the approximate educational level required to understand the text: 1.0–4.9 elementary school, 5.0–6.9 middle school (grades 5–6), 7.0–8.9 middle school (grades 7–8), 9.0–10.9 high school (grades 9–10), 11.0–12.9 high school (grades 11–12), 13.0–15.9 university (undergraduate), and ≥16.0 postgraduate/professional level.

## Statistical Analysis

Statistical analyses were performed using SPSS software (version 25.0). The normality of continuous variables was evaluated using visual methods (histograms and Q-Q plots) and the Shapiro-Wilk test. Continuous variables are presented as mean (minimum–maximum) values. Comparisons of AIRES total and subscale scores, PEMAT-P Understandability and Actionability scores, and readability indices (Flesch Reading Ease and Flesch-Kincaid Grade Level) among the three chatbots were performed using the Kruskal-Wallis test. Intra-observer and inter-observer reliability for AIRES scoring were assessed using the intraclass correlation coefficient (ICC). A p value <0.05 was considered statistically significant.

**Table 1.** Frequently asked questions on the Perthes Kids Foundation official website

| QUESTION |                                                                         |
|----------|-------------------------------------------------------------------------|
| Q1       | Is Perthes disease hereditary?                                          |
| Q2       | How common is Perthes?                                                  |
| Q3       | Is there a cure for Perthes?                                            |
| Q4       | What are the symptoms for Perthes?                                      |
| Q5       | How is Perthes disease treated?                                         |
| Q6       | How long does Perthes last?                                             |
| Q7       | Can I claim disability for Perthes disease?                             |
| Q8       | Can Perthes come back?                                                  |
| Q9       | What are some ways to help with pain management?                        |
| Q10      | What are some alternative activities my child can do for fun?           |
| Q11      | Do you have a fact sheet I can print out and give to my child's school? |

## Results

In the AIRES assessment, a statistically significant, moderate level of interobserver agreement was identified (intraclass correlation coefficient [ICC] = 0.501,  $p < 0.001$ ).

According to total AIRES scores, the responses of all three chatbots to questions related to Perthes disease were evaluated as being of excellent quality. The mean total AIRES scores were 18.63 for ChatGPT, 18.74 for Gemini, and 19.55 for DeepSeek. Detailed results for the AIRES subdomains are presented in Table 2.

**Table 2.** AIRES scores of AI chatbot responses

| AIRES             | ChatGPT<br>mean<br>(min-<br>max) | Gemini<br>mean<br>(min-<br>max) | Deepseek<br>mean<br>(min-<br>max) | p*<br>p=0.72<br>4 |
|-------------------|----------------------------------|---------------------------------|-----------------------------------|-------------------|
| Accuracy          | 4.81 (4–5)                       | 4.73 (4–5)                      | 5.00 (5–5)                        |                   |
| Comprehensiveness | 4.27 (3–5)                       | 4.64 (4–5)                      | 4.73 (3–5)                        |                   |
| Understandability | 4.82 (4–5)                       | 4.55 (3–5)                      | 4.82 (4–5)                        |                   |
| Up-to-dateness    | 4.73 (3–5)                       | 4.82 (4–5)                      | 5.00 (5–5)                        |                   |
| Mean              | 4.66                             | 4.69                            | 4.89                              |                   |

\*: Kruskal- Wallis Test

ChatGPT responses were evaluated as having sufficient understandability for all questions according to the PEMAT-P Understandability criterion. However, two responses (18.18%) were found to be insufficient in terms of Actionability.

For Gemini, one response (9.09%) was evaluated as having insufficient understandability in the PEMAT-P Understandability assessment, while four responses (36.3%) were found to have insufficient actionability in the Actionability assessment.

DeepSeek responses were evaluated as sufficient for all questions according to both the PEMAT-P Understandability and Actionability criteria.

When the PEMAT-P scores of the three chatbots were compared, Gemini was found to have statistically significantly lower Understandability scores compared with ChatGPT and DeepSeek. In contrast, no statistically significant difference was observed among the chatbots in terms of PEMAT-P Actionability scores. Detailed results are presented in Table 3.

**Table 3.** PEMAT-P Understandability (P.u) and PEMAT-P Actionability (P.a) scores of AI chatbots' responses

| PE<br>MA<br>T-P | ChatGPT  |          | Gemini   |           | Deepseek |          | p*           |                |
|-----------------|----------|----------|----------|-----------|----------|----------|--------------|----------------|
|                 | P.u      | P.a      | P.u      | P.a       | P.u      | P.a      | P.u          | P.a            |
|                 |          |          |          |           |          |          | p =<br>0.005 | p=<br>0.7<br>2 |
| q1              | 93       | 67       | 88       | 25        | 100      | 80       |              |                |
| q2              | 93       | 67       | 69       | 25        | 100      | 89       |              |                |
| q3              | 93       | 83       | 88       | 55        | 100      | 80       |              |                |
| q4              | 93       | 83       | 94       | 100       | 100      | 80       |              |                |
| q5              | 10       | 94       | 75       | 88        | 100      | 80       |              |                |
| q6              | 10       | 89       | 75       | 50        | 100      | 80       |              |                |
| q7              | 10       | 89       | 75       | 100       | 100      | 80       |              |                |
| q8              | 10       | 89       | 88       | 88        | 100      | 80       |              |                |
| q9              | 10       | 89       | 100      | 100       | 100      | 80       |              |                |
| q10             | 10       | 89       | 100      | 100       | 100      | 100      |              |                |
| q11             | 10       | 89       | 100      | 100       | 100      | 100      |              |                |
| Me<br>an        | 97.<br>4 | 84.<br>3 | 85.<br>9 | 75.<br>09 | 100      | 84.<br>4 |              |                |

\*: Kruskal- Wallis Test

ChatGPT responses were evaluated as having a "fairly difficult" readability level for all questions (100%) according to Flesch Reading Ease (FRE) scores. In the Flesch-Kincaid Grade Level (FKGL) analysis, ChatGPT was found to produce texts at a high school (grades 9–10) level for all responses.

In the FRE assessment of Gemini responses, 6 responses (54.5%) were classified as "difficult" and 5 responses (45.4%) as "fairly difficult" in terms of readability. According to the FKGL analysis, Gemini provided responses at a middle school (grades 7–8) level for 1 question (9.09%), at a high school (grades 9–10) level for 4 questions (36.3%), and at a high school (grades 11–12) level for 6 questions (54.5%).

DeepSeek responses were evaluated based on FRE scores as "difficult" for 7 responses (63.6%) and "fairly difficult" for 4 responses (36.3%). According to FKGL results, DeepSeek produced responses at a high school (grades 9–10) level for 5 questions (45.4%) and at a high school (grades 11–12) level for 6 questions (54.5%).

When the responses of the three chatbots were compared, no statistically significant difference was observed in terms of Flesch Reading Ease or Flesch-Kincaid Grade Level scores. Detailed data related to the readability analyses are presented in Table 4.

**Table 4.** Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) scores of AI chatbot responses

|      | ChatGPT   |      | Gemini    |       | Deepseek  |       | p*                |                     |
|------|-----------|------|-----------|-------|-----------|-------|-------------------|---------------------|
|      | FR<br>E   | FKGL | FRE       | FKGL  | FRE       | FKGL  | FRE<br>p=0.<br>83 | FKGL<br>p=0.13<br>3 |
| q1   | 54.<br>9  | 10.1 | 38.4      | 12.6  | 45.9      | 11.6  |                   |                     |
| q2   | 52.<br>2  | 10.3 | 52.2      | 9.6   | 47.6      | 11.5  |                   |                     |
| q3   | 52.<br>2  | 10.1 | 46.8      | 11.2  | 47.1      | 11.4  |                   |                     |
| q4   | 55.<br>0  | 10.3 | 55.6      | 9.6   | 50.4      | 10.2  |                   |                     |
| q5   | 52.<br>2  | 9.9  | 45.2      | 11.4  | 46.8      | 11.2  |                   |                     |
| q6   | 52.<br>2  | 10.1 | 49.8      | 11.1  | 49.6      | 11.1  |                   |                     |
| q7   | 52.<br>4  | 10.2 | 39.5      | 11.9  | 49.4      | 10.3  |                   |                     |
| q8   | 52.<br>4  | 10.3 | 52.2      | 10.1  | 49.8      | 11.1  |                   |                     |
| q9   | 55.<br>6  | 9.1  | 47.4      | 11.1  | 52.2      | 9.8   |                   |                     |
| q10  | 55.<br>9  | 9.2  | 55.9      | 9.4   | 52.6      | 9.5   |                   |                     |
| q11  | 55.<br>1  | 9.1  | 55.0      | 8.8   | 53.8      | 9.5   |                   |                     |
| mean | 53.<br>65 | 9.88 | 48.9<br>1 | 10.62 | 49.5<br>6 | 10.65 |                   |                     |

## Discussion

The most important finding of our study is that all evaluated chatbots were able to generate accurate, comprehensive, and up-to-date responses to questions related to Perthes disease. However, Gemini was observed to provide responses with a lower level of understandability compared with the other two chatbots, as reflected by its PEMAT-P scores. In terms of readability analyses, no difference was identified among the three chatbots, and the majority of responses were found to correspond to a high school-level readability. In conditions such as Perthes disease, which have a prolonged clinical course and an unpredictable prognosis, access to online health information may increase information-seeking behavior among patients and their families (Nassiri et al., 2015; Treadgold et al., 2020). However, numerous previous studies have demonstrated that online information in the field of pediatric orthopedics is largely of low quality (Oztermeli & Karahan, 2020; Ranade et al., 2020). Although information presented on websites with HONcode certification or on accredited and reliable sources within video-sharing platforms has been reported to be of higher quality, it has also been shown that the more sophisticated and technical content provided by these sources is less frequently preferred by users (Jeong et al., 2022; Kusec et al., 2003). This situation suggests that users may tend to favor alternative information sources that prioritize understandability and readability over strict accuracy when accessing online health information.

In this context, the questions evaluated in the present study reflect the holistic information needs encountered in pediatric orthopedic practice and are important in terms of representing real patient and parental concerns. The questions addressed not only medical information but also the social dimensions of the disease, including social activities (question 10) and school adaptation (question 11). Conventional online patient education materials often provide fragmented or insufficient responses to such multidimensional questions (Fegghi et al., 2014; Nassur et al., 2023). Therefore, the question set allowed for the evaluation of chatbots' capacity to generate responses not only related to medical treatment but also to the functional and psychosocial effects of the disease.

The high AIRES scores obtained suggest that these systems may potentially assume an advantageous role in meeting the information needs of patients and their families.

In recent years, the widespread use of chatbots has led to a significant transformation in platforms used to access health information. The ability of users to ask questions directly and to deepen information through follow-up inquiries distinguishes these systems from traditional written or visual online content (Clark & Bailey, 2024; Nasra et al., 2025). Although some studies have reported that early versions of chatbots were capable of generating inaccurate or insufficient medical responses, an increasing number of studies indicate that newer models can produce higher-quality and more reliable answers (Goodman et al., 2023). Contemporary chatbots do not present information randomly; instead, they organize sources hierarchically according to scientific authority and subject these data to consistency checks through cross-validation methods. Information that passes medical ethics and safety filters is combined with real-time literature searches via retrieval-augmented generation (RAG) technology and presented to the user within an evidence-based and up-to-date response framework (Son et al., 2025). This approach suggests that chatbots may be considered complementary tools in providing reliable health information to patients and families by maintaining accuracy while presenting information in a more understandable manner.

In our study, readability analyses of chatbot responses demonstrated that the texts were generally classified as “difficult” or “fairly difficult” according to the FRE index, while FKGL scores corresponded to a high school level.

When parents of children with Perthes disease seek online information related to treatment and social issues, responses requiring higher educational levels according to FKGL may lead to misunderstandings (Verduzco Guillen et al., 2025). Previous studies evaluating online medical information have criticized that even information presented by accredited sources is often difficult to read and have emphasized the need for more understandable content (Jeong et al., 2022; Kusec et al., 2003). The interactive nature of chatbots may allow users to improve the readability of responses by asking additional questions, thereby enabling parents to more easily comprehend complex medical and social information and potentially enhancing the effectiveness of online health education.

Future studies may evaluate the role of chatbots in pediatric orthopedics beyond the provision of static information by assessing the accuracy, understandability, and impact on clinical decision-making of responses generated when integrated with patient-specific clinical data (such as radiological stage or range of motion). In addition, personalized responses adapted to users’ reading levels, social contexts, and local healthcare practices may be anticipated to improve health literacy and treatment adherence, particularly in long-term and socially complex conditions such as Perthes disease. Ultimately, the safe integration of artificial intelligence into pediatric orthopedic practice will be shaped by prospective clinical studies determining how these tools may transform physician–patient communication within clinical decision-support processes.

The present study has several limitations. First, a limited number of questions were evaluated using only the “Frequently Asked Questions” section of the Perthes Kids Foundation website; therefore, questions reflecting broader or different social, cultural, and linguistic contexts were not examined. In addition, responses were analyzed only in the English language, and the accuracy and readability of responses in other languages may differ. Each question was posed only once, and potential variations in chatbot responses to repeated or iterative queries were not assessed. The evaluation was based on chatbot versions available at a specific time point; due to rapid system updates, results may change in future versions.

Finally, readability analyses (FRE and FKGL) provide only text-level measurements and do not directly reflect actual parental comprehension. Although these factors limit the generalizability of the findings, the results demonstrate the potential of chatbots to provide accurate, comprehensive, and readable information within the context of pediatric orthopedics and offer valuable insights for the field.

## Conclusion

This study demonstrates that contemporary chatbots are capable of generating accurate, comprehensive, up-to-date, and understandable responses to questions related to Perthes disease. Owing to their interactive structure, these systems differ from traditional online health information platforms by responding more effectively to the multidimensional information needs of patients and their families. In the future, integrating chatbots with clinical findings may further strengthen personalized information delivery and has the potential to reshape physician–patient communication in pediatric orthopedic practice.

## Ethical approval

This study does not involve any human or animal participants, and no personal or confidential patient data were utilized. The research was conducted using publicly available, anonymous data generated by artificial intelligence platforms. Therefore, institutional review board (ethics committee) approval was not required.

## Informed consent

Not applicable.

## Conflicts of interest

The authors declare that they have no conflicts of interest.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## References

- Clark, M., & Bailey, S. (2024). Chatbots in health care: Connecting patients to information (Report No. EH0122). Canadian Agency for Drugs and Technologies in Health.
- Feghhi, D. P., Komlos, D., Agarwal, N., & Sabharwal, S. (2014). Quality of online pediatric orthopaedic education materials. *The Journal of Bone and Joint Surgery*, 96(23), e194.
- Gilbert, S. R., Laine, J. C., Martin, B. D., Sankar, W. N., & Kim, H. K. W. (2025). Legg-Calvé-Perthes disease. *Journal of the American Academy of Orthopaedic Surgeons*. Advance online publication.
- Goodman, R. S., Patrinely, J. R., Stone, C. A., Jr., Zimmerman, E., Donald, R. R., Chang, S. S., ... & others. (2023). Accuracy and reliability of chatbot responses to physician questions. *JAMA Network Open*, 6(10), e2336483.
- Jeong, K. Y., Lee, H. J., & Yim, S. Y. (2022). The usefulness, reliability, and quality of YouTube video clips on congenital muscular torticollis: A STROBE-compliant study. *Medicine*, 101(37), e30502.
- Kubb, C., & Foran, H. M. (2020). Online health information seeking by parents for their children: Systematic review and agenda for further research. *Journal of Medical Internet Research*, 22(8), e19985.
- Kusec, S., Brborovic, O., & Schillinger, D. (2003). Diabetes websites accredited by the Health On the Net Foundation Code of Conduct: Readable or not? *Studies in Health Technology and Informatics*, 95, 655-660.
- Leo, D. G., Murphy, R., Gambling, T., Long, A., Jones, H., & Perry, D. C. (2019). Perspectives on the social, physical, and emotional impact of living with Perthes’ disease in children and their family: A mixed methods study. *Global Pediatric Health*, 6, 2333794X19835235.

Lock, A. M., & Baker, J. F. (2022). Quality of YouTube videos for three common pediatric hip conditions: Developmental hip dysplasia, slipped capital femoral epiphysis and Legg-Calvé-Perthes disease. *Journal of Pediatric Orthopaedics B*, 31(6), 546-553.

Nassiri, M., Bruce-Brand, R. A., O'Neill, F., Chenouri, S., & Curtin, P. (2015). Perthes disease: The quality and reliability of information on the internet. *Journal of Pediatric Orthopaedics*, 35(5), 530-535.

Nassur, J. A., Samuel, L. T., Acuña, A. J., Ellsworth, B., & Kamath, A. F. (2023). The quality of online information on Legg-Calvé-Perthes disease: Can we do better? *Archives of Orthopaedic and Trauma Surgery*, 143(11), 6569-6576.

Nasra, M., Jaffri, R., Pavlin-Premrl, D., Kok, H. K., Khabaza, A., Barras, C., ... & others. (2025). Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. *Internal Medicine Journal*, 55(1), 20-34.

Oztermeli, A., & Karahan, N. (2020). Evaluation of YouTube video content about developmental dysplasia of the hip. *Cureus*, 12(8), e9557.

Pan, A., Musheyev, D., Bockelman, D., Loeb, S., & Kabarriti, A. E. (2023). Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncology*, 9(10), 1437-1440.

Ranade, A. S., Belthur, M. V., Oka, G. A., & Malone, J. D. (2020). YouTube as an information source for clubfoot: A quality analysis of video content. *Journal of Pediatric Orthopaedics B*, 29(4), 375-378.

Son, N., Kang, I., Kim, I., Lee, K., Nam, S., & Lee, D. (2025). Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system. *Healthcare Informatics Research*, 31(3), 218-225.

Treadgold, B. M., Teasdale, E., Muller, I., Roberts, A., Coulson, N., & Santer, M. (2020). Parents' and carers' experiences of seeking health information and support online for long-term physical childhood conditions: A systematic review and thematic synthesis of qualitative research. *BMJ Open*, 10(12), e042139.

Verduzco Guillen, R., Grellinger, E. D., & Swarup, I. (2025). The availability and readability of multilingual online patient materials for Legg-Calvé-Perthes disease. *Journal of the Pediatric Orthopaedic Society of North America*, 11, 100169.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ*, 335(7624), 806-808.

Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, B. P., Raptis, D. A., ... & others. (2023). Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*, 25, e47479.

Winship, B., Grisell, M., Yang, C. B., Chen, R. X., & Bauer, A. S. (2014). The quality of pediatric orthopaedic information on the internet. *Journal of Pediatric Orthopaedics*, 34(4), 474-477.

Yun, H. S., & Bickmore, T. (2025). Online health information-seeking in the era of large language models: Cross-sectional web-based survey study. *Journal of Medical Internet Research*, 27, e68560.