BMC Musculoskeletal Disorders

RESEARCH

Open Access

High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease

Zeynel Mert Asfuroğlu[1*], Hilal Yağar[2] and Ender Gümüşoğlu[3]

Abstract

Background This study aimed to assess the quality and readability of large language model–generated responses to frequently asked questions (FAQs) about Kienböck's disease (KD).

Methods Nineteen FAQs about KD were selected, and the questions were divided into three categories: general knowledge, diagnosis, and treatment. The questions were inputted into the Chat Generative Pre-trained Transformer 4 (ChatGPT4) webpage using the zero-shot prompting method, and the responses were recorded. Hand surgeons with at least 5 years of experience and advanced English proficiency were individually contacted over instant WhatsApp messaging and requested to assess the responses. The quality of each response was analyzed by 33 experienced hand surgeons using the Global Quality Scale (GQS). The readability was assessed with the Flesch–Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES).

Results The mean GQS score was 4.28 out of a maximum of 5 points. Most raters assessed the quality as good (270 of 627 responses; 43.1%) or excellent (260 of 627 responses; 41.5%). The mean FKGL was 15.5, and the mean FRES was 23.4, both of which are considered above the college graduate level. No statistically significant differences were found in the quality and readability of responses provided for questions related to general knowledge, diagnosis, and treatment.

Conclusions ChatGPT-4 provided high-quality responses to FAQs about KD. However, the primary drawback was the poor readability of these responses. By improving the readability of ChatGPT's output, we can transform it into a valuable information resource for individuals with KD.

Level of evidence Level IV, Observational study.

Keywords Artificial intelligence, Kienböck's disease, Patient education, Readability

*Correspondence:
Zeynel Mert Asfuroğlu
z.mert.asfuroglu@gmail.com
[1]School of Medicine, Department of Orthopaedics and Traumatology, Division of Hand Surgery, University of Mersin, Mersin 33110, Turkey
[2]School of Medicine, Department of Orthopedics and Traumatology, Ömer Halisdemir University, Niğde, Turkey
[3]School of Medicine, Department of Orthopaedics and Traumatology, University of Mersin, Mersin, Turkey

Introduction

Patients most frequently use the internet for initial research related to their health information, considering the increases in technology usability and accessibility [1, 2]. More than half of all patients perform pre-consultation research on their medical conditions [3]. Therefore, access to high-quality, readable resources on the Internet is important for patients. Large language models (LLMs) are sophisticated natural language processing (NLP) systems that evaluate textual inputs to produce contextually pertinent outputs. LLMs can be defined as "Generative Artificial Intelligence (AI)," a category encompassing AI systems producing data, including text, video, or audio [4]. AI-powered chatbots, such as Chat Generative Pre-trained Transformer (ChatGPT), have recently shown significant potential in the context of improved patient education [5].

ChatGPT (OpenAI, Inc., CA, USA) is an LLM online chatbot released in November 2022. It has recently gained notable popularity, marking a significant milestone in consumer-facing deep learning. ChatGPT can grasp the complexities inherent in human language, enabling the generation of relevant, contextually appropriate responses to a wide range of questions [6]. LLM-based chatbots may play a role in providing accurate and comprehensible health information to patients in the near future, potentially reducing the spread of health misinformation and improving patient literacy [6, 7].

However, the accuracy of LLM-generated responses to health-related patient inquiries remains a significant concern [6]. Previous studies in orthopedics and traumatology were cautiously optimistic regarding the accuracy and quality of ChatGPT responses [8, 9]. However, a few studies have assessed the responses of ChatGPT in the specific field of hand surgery [10]. The increasing use of ChatGPT necessitates further examination of the accuracy, reliability, and accessibility of LLM responses to patient inquiries about hand surgery.

Kienböck's disease (KD) is characterized by avascular necrosis of the lunate carpal bone, also known as lunatomalacia [11, 12]. The Office of Rare Diseases of the National Institutes of Health classifies KD as a rare disease [13]. Obtaining accurate information on uncommon diseases such as KD is more challenging than when a disorder is common. Therefore, we investigated the quality and readability of LLM-generated responses to frequently asked questions (FAQs) about KD. We hypothesized that the responses would exhibit high quality but low readability.

Methods

Identification of frequently asked questions about Kienböck's Disease

The ethical committee approval was waived for this study because it solely comprised online information. Hence, an IRB number or informed consent form was not required. The study was conducted in accordance with the principles of the Declaration of Helsinki.

The phrases "Kienböck's Disease" and "Lunate Avascular Necrosis" were selected to represent the vocabulary commonly used by individuals with KD when searching for information about their diagnosis. Each of the two phrases was entered into the three most commonly used search engines (Google, Bing, and Yahoo). Next, FAQs about KD were identified from various websites of trustworthy organizations (my.clevelandclinic.org, www.msdmanuals.com, www.assh.org, and www.physiopedia.com). In addition, the study by Dias and Lunn was referred to develop the questions [14].

Initially, 46 FAQs about KD were selected, excluding questions with similar or vague meanings and nonmedical questions about KD. Ultimately, the authors selected 19 questions considered the most relevant by consensus. These questions were divided into three categories: general knowledge, diagnosis, and treatment. The authors modified the phrasing and grammar of several questions to ensure understandability.

The prepared questions were presented to ChatGPT-4 (OpenAI Global LLC, CA, USA, a subsidiary of OpenAI, Inc.) on March 1, 2024, with a new session for each question. The ChatGPT interface was accessed via the webpage, and the zero-shot prompting method was used.

Quality analysis

The Global Quality Scale (GQS) was used to determine the overall quality of the responses. The GQS is a 5-point Likert scale used to assess the quality of information, flow of information available online, and convenience of use. On this scale, 1 point indicates very bad quality and 5 points indicate excellent quality [15].

Hand surgeons with a minimum of 5 years of experience and advanced English language proficiency were individually contacted through instant messaging (WhatsApp; Meta Inc., CA, USA) and asked to assess the responses. Thirty-three hand surgeons analyzed the quality of each response in an online survey that included the questions and the corresponding responses provided by ChatGPT. The inter-rater reliability was also analyzed.

Readability analysis

The Flesch–Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES) were used to determine readability. The FRES ranges from 0 (indicating unreadable text) to 100 (indicating very easy-to-read text). The FKGL

translates the FRES into a corresponding reading level anticipated for a specific grade level in the American education system, ranging from first grade to college (Table S1). Both FKGL and FRES are computed using formulas that consider the average sentence length (i.e., the ratio of words to sentences) and the average number of syllables per word (i.e., the ratio of syllables to words) [16].

Statistical analysis

Statistical analysis was performed using the Statistical Package for the Social Sciences version 26.0 (IBM Corporation, NY, USA). The descriptive data were expressed as the mean ± standard deviation, or number and frequency, where applicable. After confirming the normality of all relevant variables, the *t* test was employed for independent pairwise group comparisons, and one-way analysis of variance and multivariate analysis of variance were used for comparisons involving more than two groups. The Dunn's test was used for *post hoc* analysis. Fleiss' kappa (κ) was used to assess inter-rater reliability. A P value ≤ 0.05 indicated a statistically significant difference.

Results

The compilation of "FAQs about KD", as outlined in the method section, is shown in Table 1.

The mean GQS score was 4.28 ± 0.11 (range 4.03–4.42) out of a maximum of 5 points (Fig. 1). Among the 33 raters, the most common GQS scores were good (270 of 627 responses; 43.1%) and excellent (260 of 627 responses; 41.5%). The inter-rater reliability was $\kappa = 0.74$, indicating substantial agreement.

For all responses, the mean FKGL was 15.5 ± 1.64 (range 13–20), and the FRES was 23.4 ± 9.28 (range 8–40) (Fig. 1; Table 2). The majority of the responses were considered to be above college level and extremely confusing (Table 3).

No significant differences were found in the quality or readability of responses to questions regarding general knowledge, diagnosis, and treatment (Table 2).

Discussion

The key findings of this study are that responses to FAQs about Kienböck's disease generated by ChatGPT-4 demonstrate high quality, but limited readability.

LLMs are sophisticated NLP systems that evaluate textual inputs to provide contextually pertinent outputs. LLMs have rapidly become integrated into the healthcare industry, where they are used to analyze medical images, perform robotic procedures, and support clinical decision-making [17, 18]. Considering the challenges of obtaining reliable information on the Internet, the rapid advancement of LLM is anticipated to help patients access accurate information regarding their medical conditions [19]. However, LLM-generated information has certain limitations, including inadequate knowledge of specific areas and an inability to understand context [20]. The application of ChatGPT in the field of hand surgery continues to evolve. Hence, hand surgeons should be aware of widely accessible LLMs such as ChatGPT and their capabilities to better address patients' concerns [10].

ChatGPT-4 was selected in this study because it is accessible to patients and has become a preferred

Table 1 Summary of frequently asked questions about Kienböck's disease presented to ChatGPT

Number of Question	Questions	Categories	GQS (mean ± SD)	FKGL	FRES
1	What is Kienböck's disease?	General Knowledge	4.39 ± 0.66	13.4	39.1
2	How many people have Kienböck's disease?	General Knowledge	4.24 ± 0.75	16	14.3
3	What are the risk factors for Kienböck's disease?	General Knowledge	4.42 ± 0.71	14.8	28.6
4	Is Kienböck's disease chronic?	General Knowledge	4.15 ± 0.62	15.5	19.9
5	Is Kienböck's disease bilateral?	General Knowledge	4.15 ± 0.83	17.9	10.7
6	Is Kienböck's disease work related?	General Knowledge	4.36 ± 0.65	19.4	9.6
7	How fast does Kienböck's disease progress?	General Knowledge	4.30 ± 0.64	15.1	25.9
8	What are the most recent advances in Kienböck's disease?	General Knowledge	4.30 ± 0.85	17.7	8.6
9	How painful is Kienböck's disease?	Diagnosis	4.42 ± 0.71	15.3	29.6
10	Is there any special test for Kienböck's disease?	Diagnosis	4.30 ± 0.73	13.6	35.5
11	What is Kienböck's disease Magnetic Resonance Imaging?	Diagnosis	4.27 ± 0.72	16.8	19.4
12	How do you relieve Kienböck's disease pain?	Treatment	4.30 ± 0.68	15.3	19.2
13	How long does it take to recover from Kienböck's disease?	Treatment	4.12 ± 0.82	15	25.6
14	Can you recover from Kienböck's disease?	Treatment	4.18 ± 0.73	16.6	19.9
15	What happens if Kienböck's disease is left untreated?	Treatment	4.30 ± 0.77	13.5	31
16	How do you treat Kienböck's disease without surgery?	Treatment	4.36 ± 0.55	15.1	22.5
17	How long does Kienböck's disease surgery take?	Treatment	4.03 ± 0.64	13.5	35.1
18	What is the success rate of Kienböck's disease surgery?	Treatment	4.33 ± 0.65	16.5	16.1
19	What is the best surgery for Kienböck's disease?	Treatment	4.27 ± 0.76	14.2	34.5

GQS: Global quality scale, FKGL: Flesch-Kincaid grade level, FRES: Flesch reading ease score, SD: Standard deviation

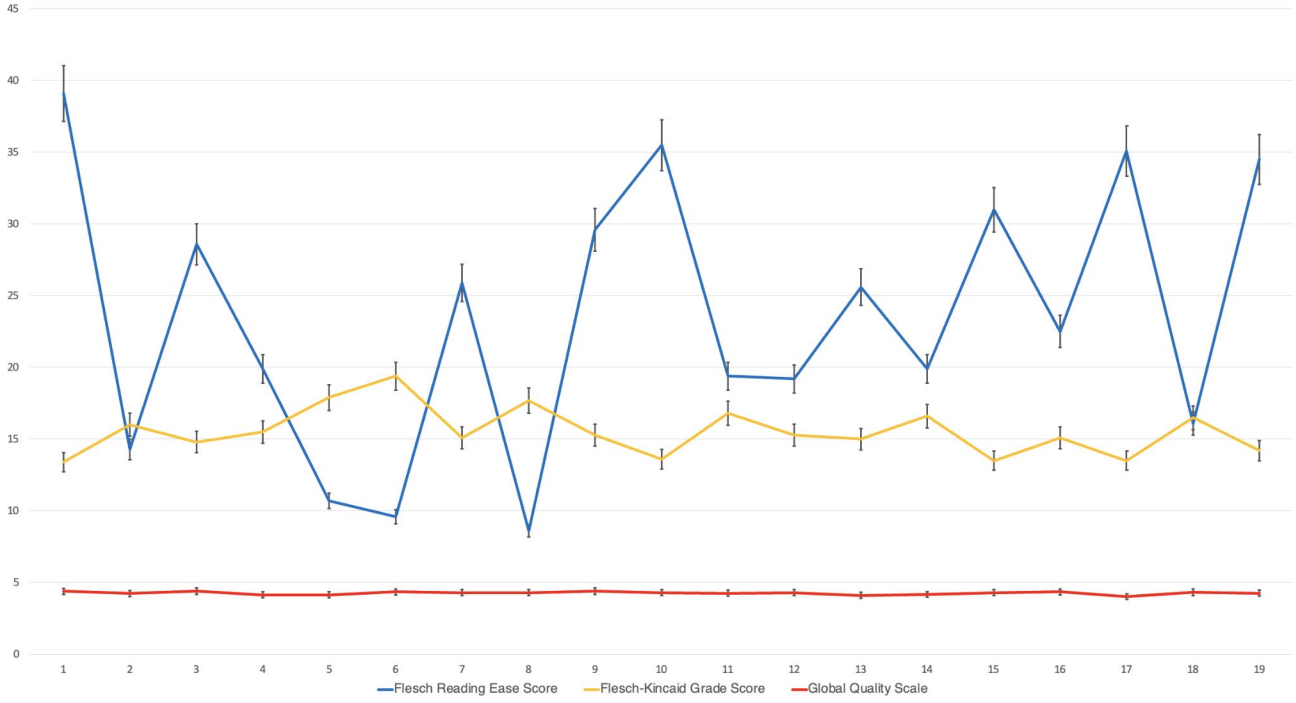

Fig. 1 Line graphs showing the FRES (blue line), FKGL (yellow line) and GQS (red line) scores. The *x*-axis represents the score, and the *y*-axis represents the questions

Table 2 Quality and readability statistics stratified by the question categories

	General knowledge	Diagnosis	Treatment	Total	*p*-value
GQS (mean ± SD)	4.25 ± 0.1	4.29 ± 0.02	4.22 ± 0.12	4.28 ± 0.11	0.615
FKGL (mean ± SD)	16.22 ± 1.82	15.23 ± 1.02	14.96 ± 1.36	15.5 ± 1.64	0.498
FRES (mean ± SD)	19.58 ± 11.04	28.16 ± 8.83	25.48 ± 6.81	23.4 ± 9.28	0.28

SD: standard deviation

Table 3 Categorical readability results of the Flesch reading ease score (FRES) and Flesch-Kincaid grade level (FKGL)

Test	*n* (%)	Level of readability
Flesch reading ease score	5 (26.3%)	College level
	14 (73.7%)	Above college level
Flesch-Kincaid grading	6 (31.5%)	Difficult
	13 (68.5%)	Extremely confusing

n: number

instrument in the chatbot arena. An important limitation of ChatGPT is its propensity to incorporate overt inaccuracies in its responses, commonly termed "artificial hallucination" [21]. Proper prompting techniques can produce outputs closely corresponding with pertinent data and established knowledge, thereby diminishing the probability of hallucination. Clear and concise prompts reduce ambiguity, thus enhancing the precision of information [22]. In this study, special attention was paid to ensure that the prepared questions were clear and concise, and the questions were inputted into the ChatGPT-4 using the zero-shot prompting method. Zero-shot prompting refers to performing tasks without any specific examples [23]. The reason for choosing this method was to assess the level of ChatGPT-4's pre-existing knowledge without any prior information.

The GQS was used as a quality assessment tool to enable more raters to participate using a relatively simple assessment method. In our study, 33 experienced hand surgeons, each with a minimum of 5 years of experience, evaluated the responses. The high inter-rater agreement indicates rater homogeneity and assessment reliability. Readability was evaluated using the FKGL and FRES, which are widely used in the literature and have proven to be reliable [1, 10, 24].

Crook et al. [10]. analyzed the information obtained by ChatGPT for common diseases requiring hand surgery (carpal tunnel syndrome, cubital tunnel syndrome, trigger finger, and distal radius fracture) and reported that the responses exhibited high quality. Numerous studies across various medical specialties have analyzed AI-generated responses [25–30]. These studies indicate that AI-generated responses exhibit good quality for medical disciplines other than hand surgery. Taşkaldıran et al. [27] reported that the mean GQS score was 4.9 for

ChatGPT-4 responses regarding hyperparathyroidism. Our findings were consistent with the previous studies in this regard. The 33 experienced hand surgeons rated the responses as high quality, independent of the question category (Table 3).

For written text to be comprehensible, it must be readable. The use of numerous unfamiliar terms, frequent repetition of these words, and a high proportion of passive sentences all contribute to diminished text readability [16, 31]. The abundance of medical terminology in health-related informative text reduces the readability. Hadden et al. [32]. assessed the readability of educational materials on hand surgery. The average Flesch–Kincaid grade of texts on various disorders was 9.3 (difficult to read). Our results showed that the categorical readability levels were considered to be above college level and extremely confusing. Therefore, the main disadvantage of the AI-generated responses was poor readability (Table 3), leading to doubts about the current feasibility and usability of ChatGPT by the general population.

This study is novel in assessing AI-generated responses to FAQs about KD. A few studies have assessed the online information sources related to KD. Noback et al. [24]. highlighted the need for health literacy to gain information about KD on the Internet due to the low readability of online resources. Katt et al. [1]. reported that some of the informative websites about KD were commercial in nature, and the provided information exhibited limited completeness. These studies indicated that the online information sources for KD were inadequate. We believe that AI-based software such as ChatGPT can make information about KD more accessible, enhancing both efficiency and convenience.

This study had some limitations. First, we used a single AI software, ChatGPT-4, and did not compare it with any other AI software or websites to increase the number of raters. The responses sent to the 33 different raters spanned 15 pages. Such a large number of raters might make assessments extremely difficult if we were to compare these responses with other AI software. The second limitation was the potential for "artificial hallucination." This study did not individually assess inaccuracies; however, we believe that the risk of "artificial hallucination" was mitigated due to the clarity of our questions and the inclusion of precise information in the prompts.

Conclusions

The LLM chatbot provided high-quality responses to FAQs about KD. However, a substantial disadvantage exists regarding the readability of the information provided. Hence, if LLMs improve their readability or if societal health literacy increases, LLM chatbots can become a valuable source of information, even for rare diseases such as KD.

Abbreviations

AI	Artificial intelligence
ChatGPT	Chat Generative Pre-trained Transformer
FAQs	Frequently asked questions
FKGL	Flesch–Kincaid Grade Level
FRES	Flesch Reading Ease Score
GQS	Global Quality Scale
KD	Kienböck's disease
LLM	Large language model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12891-024-07983-0.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

Not applicable.

Author contributions

ZMA: Conseption and design of the work, data analysis, writing / HY: data collection, data analysis / EG: data collection, data analysis.

Funding

Not applicable.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 August 2024 / Accepted: 21 October 2024
Published online: 04 November 2024

References

1. Katt BM, Lucenti L, Mubin NF, Nakashian M, Fletcher D, Aita D, et al. An evaluation of the source and content of Kienböck's Disease Information on the internet. J Hand Microsurg. 2021;13:65–8.
2. Swoboda CM, Van Hulle JM, McAlearney AS, Huerta TR. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National trends Survey (HINTS). BMC Fam Pract. 2018;19:146.
3. Rao AJ, Dy CJ, Goldfarb CA, Cohen MS, Wysocki RW. Patient preferences and utilization of online resources for patients treated in hand surgery practices. Hand (N Y). 2019;14:277–83.
4. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large Language models: a Comprehensive Roadmap for Healthcare Integration. Healthc (Basel). 2023;11(20):2776.
5. Villarreal-Espinosa JB, Berreta RS, Allende F, Garcia JR, Ayala S, Familiari F, et al. Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery. Knee. 2024;51:84–92.
6. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.

7. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568.
8. Subramanian T, Shahi P, Araghi K, Mayaan O, Amen TB, Iyer S, et al. Using artificial intelligence to answer common patient-focused questions in minimally invasive spine surgery. J Bone Joint Surg Am. 2023;105:1649–53.
9. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg. 2023;105:1519–26.
10. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of Online Artificial Intelligence-Generated Information on Common Hand Procedures. J Hand Surg Am. 2023;48:1122–7.
11. Camus EJ, Van Overstraeten L. Kienböck's disease in 2021. Orthop Traumatol Surg Res. 2022;108:103161.
12. Schuind F, Eslami S, Ledoux P. Kienbock's disease. J Bone Joint Surg Br. 2008;90:133–9.
13. Daly CA, Graf AR. Kienböck Disease: clinical presentation, epidemiology, and historical perspective. Hand Clin. 2022;38:385–92.
14. Dias JJ, Lunn P. Ten questions on Kienbock's disease of the lunate. J Hand Surg Eur Vol. 2010;35(7):538–43.
15. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014;96:395–403.
16. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32:221–33.
17. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in Healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47:33.
18. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8:188–94.
19. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of Medical Information provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res. 2023;25:e47479.
20. Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes E, Bauer M. Artificial intelligence and increasing misinformation. Br J Psychiatry. 2024;224(2):33–5.
21. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med. 2023;388:1233–9.
22. Shah K, Xu AY, Sharma Y, Daher M, McDonald C, Diebo BG, et al. Large Language Model Prompting techniques for Advancement in Clinical Medicine. J Clin Med. 2024;13(17):5101.
23. Meskó B. Prompt Engineering as an important emerging skill for medical professionals: Tutorial. J Med Internet Res. 2023;25:e50638.
24. Noback PC, Trofa DP, Dziesinski LK, Trupia EP, Galle S, Rosenwasser MP. Kienböck Disease: Quality, Accuracy, and readability of Online Information. Hand (N Y). 2020;15:563–72.
25. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589–96.
26. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med. 2023;19:1989–95.
27. Taşkaldıran I, Emir Önder Ç, Gökbulut P, Koç G, Kuşkonmaz ŞM. Evaluation of the accuracy and quality of ChatGPT-4 responses for hyperparathyroidism patients discussed at multidisciplinary endocrinology meetings. Digit Health. 2024;10:20552076241278692.
28. Rasmussen MLR, Larsen A-C, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefes Arch Clin Exp Ophthalmol. 2023;261:3041–3.
29. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the Language Model ChatGPT to questions regarding bariatric surgery. Obes Surg. 2023;33:1790–6.
30. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs. 2024;23:95–8.
31. Chall JS. Readibility: the beginning years. In: Zakaluk B, Samuels SJ, editors. Readability: its past, Present and Future. International Reading Association Inc. Newark; 1988. pp. 3–4.
32. Hadden K, Prince LY, Schnaekel A, Couch CG, Stephenson JM, Wyrick TO. Readability of patient education materials in hand surgery and health literacy best practices for improvement. J Hand Surg. 2016;41:825–32.

Publisher's note