DOI: 10.37990/medr.1730013 Med Records 2026;1:1730013

MEDICAL RECORDS-International Medical Journal

Research Article

Evaluating ChatGPT as a Patient Education Tool: Insights on Quality, Readability, and Reliability for Trigger Finger Frequently Asked Questions

Mustafa Çeltik[1], Ertuğrul Şahin[2], Selahaddin Aydemir[3], Yağmur Işın[4], Murat Çiçeklidağ[5], Onur Başçı[6], Mustafa Hulisi Özkan[6]

[1]Orthopedics, İzmir Bakırçay University, İzmir, Türkiye
[2]Orthopedics, Kafkas University Faculty of Medicine, Kars, Türkiye
[3]Orthopedics, Kastamonu Research and Training Hospital, Kastamonu, Türkiye
[4]Orthopedics, Tepecik Research and Training Hospital, İzmir , Türkiye
[5]Orthopedics, Siverek State Hospital, Şanlıurfa, Türkiye
[6]Orthopedics, Dokuz Eylül University Faculty of Medicine, İzmir, Türkiye

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial-NonDerivatives 4.0 International License.

Abstract

Aim: Trigger finger (TF), or stenosing tenosynovitis, causes pain, snapping, and finger locking. It greatly affects patients' quality of life, prompting frequent inquiries to healthcare providers. ChatGPT, an AI language model, has gained popularity as a tool for patient education. This study evaluated the quality and readability of ChatGPT's responses to common TF FAQs.

Material and Methods: A set of FAQs regarding TF was developed based on reputable sources such as WebMD, Mayo Clinic, and NHS Trusts. Two experienced surgeons reviewed and refined the questions before submitting them to ChatGPT-4 for response generation. The quality of the responses was evaluated using the Global Quality Score (GQS) and DISCERN scale, while readability was assessed using the Flesch Reading Ease Score (FRES) and Flesch–Kincaid Grade Level (FKGL). Inter-rater reliability was determined using Cohen's Kappa.

Results: Overall quality was generally acceptable. The mean GQS was 3.8, with mean scores ranging from 3.0 to 4.5 across questions. The mean DISCERN score was 39.83 ± 8.43 (range 30.0–53.0), indicating overall fair quality, although some responses reached the good range. Inter-rater agreement was high (Cohen's Kappa= 0.91). Readability was frequently suboptimal for patient education. The mean FRES was 39.98 (range 6.93–78.03), and only two responses met the recommended threshold of ≥60. The mean FKGL was 12.08 (range 6.43–18.30), and all responses except one exceeded the commonly recommended patient education level of grade 8.

Conclusion: ChatGPT-4 produced generally acceptable trigger finger answers by GQS, but overall reliability (DISCERN) and readability (FRES/FKGL) were often inadequate for patient education. Clinician oversight and plain-language optimization are necessary before such outputs are used.

Keywords: Trigger finger, ChatGPT, Patient education, Readability

INTRODUCTION

Trigger finger or stenosing tenosynovitis is a disease that occurs after repetitive hand use and is often characterized by pain, snapping, and locking sensations in the fingers. (1) It causes serious functional loss and decreased quality of life in patients. (2) As a result, patients have many questions about this condition and turn to health professionals and information sources to answer them.

In addition to physical therapy methods and interventional procedures (dry needling(3), steroid injections(4,5), etc.), surgical treatment interventions are also included in the treatment of trigger finger.

Due to the high number of treatment options and the diversity of the clinical presentation of the disease, information tools are of great importance in patient awareness and education. Individuals commonly turn to the internet to inquire about

CITATION
Çeltik M, Şahin E, Aydemir S, Işın Y, Çiçeklidağ M, Başçı O, Özkan MH. Evaluating ChatGPT as a Patient Education Tool: Insights on Quality, Readability, and Reliability for Trigger Finger Frequently Asked Questions. Med Rec 2026;1: 1730013. DOI: 10.37990/medr. 1730013

Received: 01.08.2025 Accepted: 03.09.2025 Published: 05.02.2026
Corresponding Author: Selahaddin Aydemir, Orthopedics, Kastamonu Research and Training Hospital, Kastamonu, Türkiye
E-mail: selahaddinaydemir@gmail.com

health-related topics, particularly given the enhanced technology accessibility (6). Routine internet sources offer great convenience in accessing information but also confront patients with the risk of disinformation.

In recent years, the Chat Generative Pre-trained Transformer (ChatGPT) (OpenAI, Inc., CA, USA), developed by OpenAI, has emerged as a widely-used tool for information sharing, particularly in enhancing patient education (7–9). For this reason, our study aimed to evaluate the answers given by ChatGPT to the most frequently asked questions about the trigger finger by patients to ChatGPT in terms of fluency, information accuracy, and level of readability for the patient. We hypothesized that the responses generated by ChatGPT would be of high quality but moderate readability.

MATERIAL AND METHODS

This study analyzed responses generated by ChatGPT to the FAQs about the “trigger finger” to provide patients with educational information (10). A list of the most frequently asked questions on the ‘trigger finger’, was identified from numerous websites of trustworthy organizations such as WebMD, Mayo Clinic, Cleveland Clinic, Medscape, and various NHS Trusts. Questions were considered for inclusion if they were repeatedly identified across more than one reliable health information source, were clearly related to trigger finger, and represented issues commonly raised by patients in clinical practice. Items were excluded when they were unclear, overlapping with other questions, excessively technical (such as detailed surgical procedures unlikely to be asked by patients), or not directly relevant to patient education. Two surgeons with more than 10 years of experience in hand and upper extremity surgery (MHO, OB) evaluated the questions to determine whether they were suitable for evaluation. Open and closed-ended questions were considered, and any ambiguous or irrelevant items were excluded during this process.

Responses were generated after the questions were finalized using ChatGPT, a text-based artificial intelligence tool. The prepared questions were presented to ChatGPT-4 (OpenAI Global LLC, CA, USA, a subsidiary of OpenAI, Inc.) on December 10, 2024. The browser cache and browsing history were cleared before response generation. The responses were evaluated for quality and readability.

The quality of the responses was assessed by two independent raters, experienced hand and upper extremity surgeons, using the Global Quality Score (GQS) (Table 1) and DISCERN scores, standardized methods to evaluate clarity, accuracy, and relevance. For GQS assessment, the rating ranges from one (poor quality) to five (excellent quality). The total DISCERN (Table 2) score can range from 16 to 80, with scores between 16 and 27 considered to be very poor, between 28 and 38 to be poor, between 39 and 50 to be fair, between 51 and 62 to be good, and between 63 and 80 to be excellent (11).

The readability was analyzed using the Flesch Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL) to ensure the language used was accessible to a broad patient demographic (Table 3). The FRES ranges from 0 (indicating unreadable text) to 100 (indicating very easy-to-read text). The FKGL converts the FRES into an expected reading level for a particular grade within the U.S. education system, from first grade to college. These readability tools help measure readability by incorporating the number of words and syllables in sentences and words (12). The readability scores were calculated by copying ChatGPT’s responses for each question to the free online readability calculator tool (13).

Ethical approval

Since ChatGPT-4 is a public application and there are no human or animal participants, ethics committee approval was not required.

Statistical Methods

All statistical analyses were performed using SPSS version 22 (IBM SPSS Statistics for Windows, Armonk, NY: IBM Corp). The agreement between the two authors who independently evaluated the ChatGPT responses was tested using Cohen’s Kappa coefficient. Shapiro–Wilk test was used to assess the normality of the distribution. Normally distributed data were presented as mean ± SD. Categorical variables were expressed as numbers and percentages.

Flesch Reading Ease (FRES) formula:

Table 1. The Global Quality Scoring System

GQS	Score
Poor quality, poor flow of the site, most information missing, not at all useful for patients	1
Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients	2
Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients	3
Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients	4
Excellent quality and excellent flow, very useful for patients	5

Table 2: Questions of DISCERN Scale

Question No	Question	Score
1	Are the aims clear?	1 2 3 4 5
2	Does it achieve its aims?	1 2 3 4 5
3	Is it relevant?	1 2 3 4 5
4	Is it clear what sources of information were used to compile the publication (other than the author)?	1 2 3 4 5
5	Is it clear when the information used or reported in the publication was produced?	1 2 3 4 5
6	Is it balanced and unbiased?	1 2 3 4 5
7	Does it provide details of additional sources of support and information?	1 2 3 4 5
8	Does it refer to areas of uncertainty?	1 2 3 4 5
9	Does it describe how each treatment works?	1 2 3 4 5
10	Does it describe the benefits of each treatment?	1 2 3 4 5
11	Does it describe the risks of each treatment?	1 2 3 4 5
12	Does it describe what would happen if no treatment is used?	1 2 3 4 5
13	Does it describe how the treatment choices affect overall quality of life?	1 2 3 4 5
14	Is it clear that there may be more than one possible treatment choice?	1 2 3 4 5
15	Does it provide support for shared decision-making?	1 2 3 4 5
16	Based on the answers to all of the above questions, rate the overall quality of the publication	1 2 3 4 5

Table 3. Readability metrics and formulas (Flesch Reading Ease and Flesch–Kincaid Grade Level)

Score	School Level (US)	Notes
100 - 90	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90 - 80	6th grade	Easy to read. Conversational English for consumers.
80 - 70	7th grade	Fairly easy to read.
70 - 60	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60 - 50	10th to 12th grade	Fairly difficult to read.
50 - 30	College	Difficult to read.
30 - 10	College Graduate	Very difficult to read. Best understood by university graduates.
10 - 0	Professional	Extremely difficult to read. Best understood by university graduates.

- FRES = 206.835 − 1.015 × (Total words / Total sentences) − 84.6 × (Total syllables / Total words)
- Flesch–Kincaid Grade Level (FKGL) formula:
- FKGL = 0.39 × (Total words / Total sentences) + 11.8 × (Total syllables / Total words) − 15.59

Higher FRES scores indicate easier-to-read text (better readability). Higher FKGL scores indicate more complex text that typically requires a higher grade/education level to understand.

RESULTS

In this study, 20 questions determined by our authors within the framework of the sources and the answers generated by ChatGPT (model: GPT-4; OpenAI, San Francisco, CA, USA) were evaluated. The evaluation was organized into specific categories to ensure a thorough assessment. The distribution of questions was organized under subheadings as follows: four questions were categorized under “General Topic,” three questions under “Symptoms and Diagnosis,” seven questions under “Treatment Options,” and six questions under “Aftercare and Prognosis”. Two experienced surgeons used the GQS and DISCERN scales to evaluate the quality of the answers generated. Readability was assessed using the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL).

Quality Assessment

The mean GQS score is 3.8. The mean Global Quality Score (GQS) for the responses ranged from 3 to 4.5. These scores indicate a generally high quality across the responses, with a standard deviation of up to 0.707, reflecting moderate variability in the raters’ evaluations. Q12 and Q4 had the highest GQS scores, while Q7 and Q17 had the lowest scores (Table 4). Responses with the highest GQS scores demonstrated superior clarity, accuracy, and patient-centered communication. In contrast, lower-scoring responses highlighted improvement areas, including simplifying medical terminology and ensuring complete topic coverage. The mean GQS scores distribution is illustrated in the bar graph, showing quality levels achieved (Figure 1).

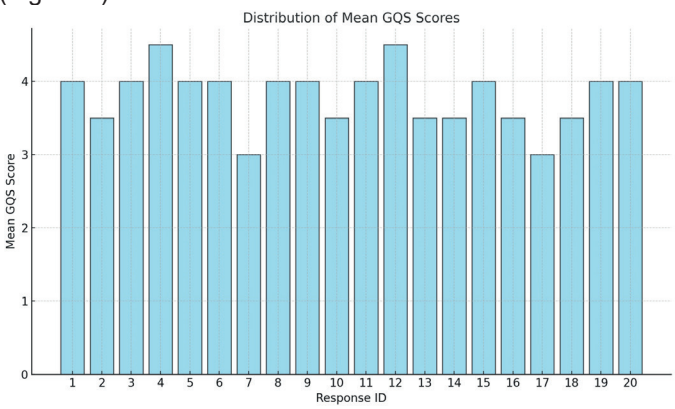

Figure 1. A bar graph summarizing the distribution of mean GQS scores

Table 4: Readability, Quality, and DISCERN Scores for Trigger Finger (TF) Questions

Category	Question	Flesch Reading Ease Score (FRES)	Flesch–Kincaid Grade Level (FKGL)	Global Quality Scale (GQS) (Mean)	DISCERN Scores (Mean)
General	Q1 What is trigger finger (TF)?	44.41	13.28	4.00	31.50
	Q2 What causes trigger finger (TF)?	27.84	13.6	3.50	31.50
	Q3 Can my trigger finger heal on its own?	39.33	11.5	4.00	30.50
	Q4 Is trigger finger the same as arthritis?	43.57	10.66	4.50	30.00
Symptoms and Diagnosis	Q5 What are the signs and symptoms of trigger finger (TF)?	35.11	18.3	4.00	31.00
	Q6 How is trigger finger diagnosed?	43.96	11.23	4.00	32.00
	Q7 Are my other fingers at risk of developing this condition?	41.59	12.55	3.00	32.50
Treatment Options	Q8 What are the treatment options available?	6.93	18.01	4.00	49.00
	Q9 What are the non-surgical treatments for trigger finger?	17.33	13.28	4.00	51.50
	Q10 Will wearing a splint help with my trigger finger?	49.09	10.51	3.50	36.50
	Q11 What medications can relieve my trigger finger pain?	27.69	12.63	4.00	40.50
	Q12 Do I need surgery for my trigger finger?	17.02	15.36	4.50	53.00
	Q13 What are the risks and possible complications from surgery?	14.93	13.91	3.50	50.50
	Q14 What happens if my trigger finger is left untreated?	52.73	10.13	3.50	51.00
Aftercare and Prognosis	Q15 How long does it take to recover from trigger finger?	61.7	8.13	4.00	50.00
	Q16 How soon after surgery can I drive?	78.03	6.43	3.50	34.00
	Q17 When can I return to work?	51.91	11.11	3.00	35.00
	Q18 Can trigger finger come back after treatment?	41.7	12.29	3.50	40.50
	Q19 Will I regain full use of my finger after surgery?	56.13	9.78	4.00	41.00
	Q20 How can I prevent trigger finger in the future?	48.64	8.96	4.00	45.00

Two raters independently assessed the DISCERN scores for 20 responses, which were analyzed to calculate DISCERN scores. The mean DISCERN score was 39.83 ± 8.43. The mean DISCERN scores for the responses ranged from 30.0 to 53.0, reflecting variations in the overall quality of the assessed responses (Table 4). The lowest average score of 30.0 was reported for Q4 whereas the highest average score of 53.0 was seen for Q12. The inter-rater reliability, determined through Cohen's Kappa, was found to be 0.91, indicating a high level of agreement between the two raters.

Readability Assessment

The readability of the responses was evaluated using two standard metrics: the Flesch Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL). These metrics assess text accessibility and complexity to determine whether the information generated suits a general patient population. The FRES rates readability on a scale of 0 to 100; the higher the score, the easier the text is to read. The average score was 39.98, with scores ranging from 6.93 to 78.03, indicating that a large portion of the material can be difficult for readers who are not college-educated. Questions with more complex structure and technical terminology—such as Q8—had the lowest FRES score (6.93). On the other hand, the highest FRES score was for Q16 ("How soon after surgery can I drive?") at 78.03, suggesting very easy readability. The FRES score of 60 marks the recommended readability threshold for health materials. (12) None of the responses except Q15 and Q16 met the 60-point readability criteria (Figure 2). FAQs addressing treatment options (Q8, Q9, Q12) were particularly challenging to read, with average FRES scores below 20. FAQs about prognosis (Q14, Q15, Q19) showed improved FRES scores, which aligns with the goal of clarifying patient recovery information (Figure 2).

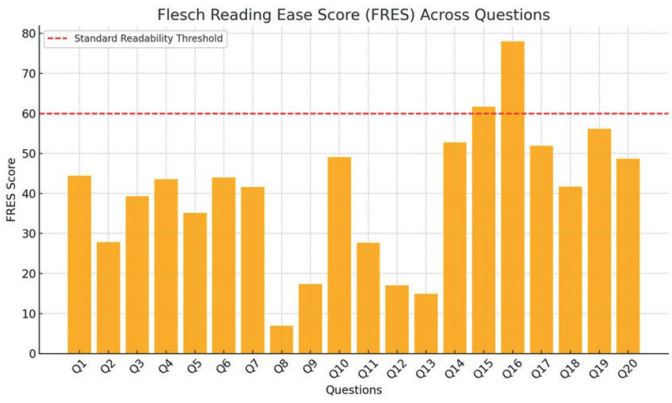

Figure 2. The red dashed line at a FRES score of 60 marks the recommended readability threshold for health materials.

The Flesch–Kincaid Grade Level (FKGL) reflects the U.S. grade level needed to understand the text. The scores from our data set ranged from 6.43 (middle-school reading level) to 18.30 (graduate-level difficulty) (Figure 3). The mean for FKGL scores was 12.08, indicating that, on average, the text requires an 11th- to 12th-grade reading level to comprehend. The highest FKGL was observed in Q5 (18.3). On the other hand, Q16 ("How soon after surgery can I drive?") had one of the lowest scores at 6.43. Apart from Q16, all the responses exceeded the optimal grade level of 8 for general patient education materials.

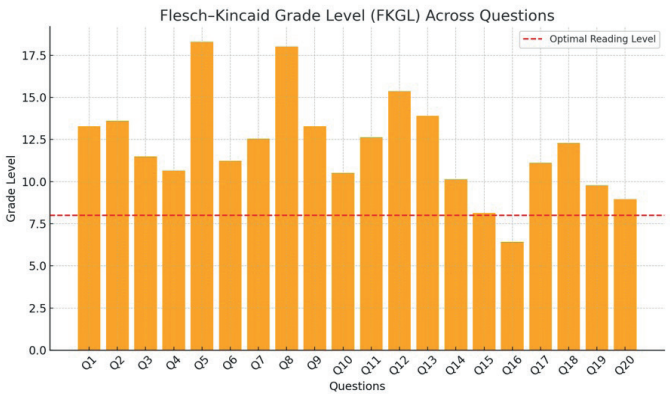

Figure 3. The red dashed line indicates an optimal grade level of 8 for patient education materials. Many treatment-related questions (e.g., Q8, Q12) exceeded this level, while recovery and prognosis questions (e.g., Q14–Q20) had lower FKGL scores, indicating better readability for patients.

DISCUSSION

This study assessed the quality and readability of responses generated by ChatGPT for patients seeking information about trigger finger. The results showed that responses from ChatGPT-4 demonstrated good reliability and quality. Regarding readability, the content typically necessitates a college-level education, suggesting it is often categorized as difficult or very difficult to read.

Patients often worry about their disease and struggle to find accurate information. Many use the internet as their primary source for health-related information (14). ChatGPT is a popular advanced AI language model and is increasingly used by patients seeking answers to their concerns and questions (15). The Global Quality Score (GQS) and DISCERN scale evaluations demonstrated that most of ChatGPT's responses were of moderate to high quality. The mean GQS was 3.8, suggesting that the answers generally provided clear, accurate, and patient-centered information. While general questions scored the highest in quality, the risk factors or post-operative guidance achieved lower scores, revealing gaps in clarity or completeness. DISCERN scores ranged from 30 to 53 (poor to good), with a mean of 39.83 ± 8.43, indicating overall fair quality. Some questions achieved scores in the good range, while others showed gaps in balance and coverage, especially around treatment risks and alternative options. This variability underscores the importance of human oversight in assessing AI-generated patient education information.

Some questions, such as Q7 and Q17, received lower GQS and DISCERN scores. These items often involved context-specific issues, for example individual risk factors or post-operative advice, where the answers were either too general or missing important details. In some cases, the use of technical terms also reduced patient clarity. Our findings indicate that ChatGPT provided clearer answers to general questions than to more context-specific ones. Future improvements may be achieved by incorporating clinical guidelines and patient education materials into AI training.

The mean FRES score was 39.98, indicating that most of the content was difficult to read for individuals without a college education. The FKGL scores supported these findings, with a mean of 12.08, suggesting that the content required a 12th-grade reading level to comprehend. These findings align with previous studies evaluating ChatGPT's performance in patient education across various medical fields. The average adult in the United States reads at about a 7th–8th grade level, and health authorities generally advise that patient information should not be written above this level. This means that a considerable number of patients may have difficulty understanding ChatGPT's current responses, making it important to use simpler language and adjust content to different literacy needs.

Numerous studies in different fields have extensively examined the reliability and readability of ChatGPT in patient education. Haidar O et al. highlighted information quality and readability issues in vascular studies, stressing the importance of patients relying on reputable, human-generated sources(16). Onder CE et al. found good information quality on hypothyroidism during pregnancy but noted poor readability(17). There are also examples of the use of ChatGPT in orthopedics and traumatology in terms of patient education. Li et al. concluded that ChatGPT provided generally accurate information to common questions about ACL reconstruction but stated that clarification often was needed regarding patient inquiries about ACL tears and ACL reconstruction (18). There are detailed studies with similar results also in hand surgery(8,19,20).

The observed gaps in clarity, comprehensiveness, and readability pointed out in the study underline the need for human oversight and content customization to ensure that AI-generated materials meet the diverse needs of patients.

Limitations

The sample size of the study is relatively small, which could restrict the generalizability of the results. Furthermore, because of this limited sample size, we did not assess outcomes across various categories. Additionally, we utilized only one AI software, ChatGPT-4, and did not compare it with other AI tools or websites to expand the number of raters involved. Lastly, while working with AI, there is also a risk of encountering artificial hallucinations.

CONCLUSION

While ChatGPT demonstrates significant potential as a patient education tool, human oversight is essential for optimizing AI tools in healthcare education regarding trigger finger. Finally, efforts should be made to improve the reliability and readability of ChatGPT and to develop orthopedics-specific models.

Acknowledgements

None to state

Funding

The authors did not receive support from any organization for the submitted work.

Competing interests

The authors declare no competing interests.

REFERENCES

1. Makkouk AH, Oetgen ME, Swigart CR, Dodds SD. Trigger finger: etiology, evaluation, and treatment. Curr Rev Musculoskelet Med. 2008;1:92–6.
2. Langer D, Maeir A, Michailevich M, Applebaum Y, Luria S. Using the international classification of functioning to examine the impact of trigger finger. Disabil Rehabil. 2016;38(26):2530–7.
3. El-halawany EF, Yamany AAER, Gaballah SAA, Atta HK. Effect of dry needling on quality of life in patients with trigger finger: a randomized controlled trial. Bull Fac Phys Ther. 2024;29(1):11.
4. Dala-Ali BM, Nakhdjevani A, Lloyd MA, Schreuder FB. The efficacy of steroid injection in the treatment of trigger finger. Clin Orthop Surg. 2012;4(4):263–8.
5. Shakeel H, Ahmad TS. Steroid injection versus NSAID injection for trigger finger: a comparative study of early outcomes. J Hand Surg Am. 2012;37(7):1319–23.
6. 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. 2019;14(2):277–83.
7. AlShenaiber A, Datta S, Mosa AJ, Binhammer PA, Ing EB. Large Language Models in the Diagnosis of Hand and Peripheral Nerve Injuries: An Evaluation of ChatGPT and the Isabel Differential Diagnosis Generator. J Hand Surg Glob Online. 2024;6(6):847–54.
8. Leypold T, Schäfer B, Boos AM, Beier JP. Artificial Intelligence-Powered Hand Surgery Consultation: GPT-4 as an Assistant in a Hand Surgery Outpatient Clinic. J Hand Surg Am. 2024;49(11):1078–88.
9. 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(1):33.
10. OpenAI TB. Chatgpt: Optimizing language models for dialogue. OpenAI. 2022.
11. Gunduz ME, Matis GK, Ozduran E, Hanci V. Evaluating the Readability, Quality, and Reliability of Online Patient Education Materials on Spinal Cord Stimulation. Turk Neurosurg. 2024;34(4).
12. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221.
13. Simpson D. The readability test tool. Access 21/10/2011. http://www. read-able. com; 2013.
14. Bw H. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. Arch Intern Med. 2005;165:2618–24.
15. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44.
16. Haidar O, Jaques A, McCaughran PW, Metcalfe MJ. AI-Generated Information for Vascular Patients: Assessing the Standard of Procedure-Specific Information Provided by the ChatGPT AI-Language Model. Cureus. 2023;15(11).
17. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 2024;14(1):243.
18. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. Arthrosc J Arthrosc Relat Surg. 2024;
19. Tan RES, Teo WZW, Puhaindran ME. Artificial Intelligence in Hand Surgery–How Generative AI is Transforming the Hand Surgery Landscape. J Hand Surg (Asian-Pacific Vol. 2024;29(02):81–7.
20. Leypold T, Schäfer B, Boos A, Beier JP. Can AI think like a plastic surgeon? Evaluating GPT-4's clinical judgment in reconstructive procedures of the upper extremity. Plast Reconstr Surgery–Global Open. 2023;11(12):e5471.