AHOT Acta Haematologica Oncologica Turcica

Acta Haematol Oncol Turc

Original Article

DOI: 10.4274/ahot.galenos.2025.2024-10-7

Readability and Quality Analysis of ChatGPT o1's Responses on Colorectal Cancer: A Study of an AI Language Model

Efe Cem Erdat, Merih Yalçıner, Güngör Utkan

Ankara University Faculty of Medicine, Department of Medical Oncology, Ankara, Türkiye

ABSTRACT

Aim: This study aimed to evaluate the readability and quality of the responses provided by the ChatGPT model to the most frequently searched questions by patients about colorectal cancer on the internet.

Methods: The 20 most frequently searched topics related to colorectal cancer were identified from Google and Yandex search engine statistics. These topics were posed to the ChatGPT o1 model, and the obtained responses were analyzed for readability using the Ateşman and Çetinkaya-Uzun readability formulas. Quality assessment was performed using the DISCERN instrument and the Global Quality Score (GQS). Statistical analyses included Pearson correlation and one-way ANOVA tests.

Results: The average word count of the responses was 654.9 [standard deviation (SD)=221.62]. According to the Ateşman readability score, the mean score was 55.45 (SD=6.06, medium difficulty readability), and according to the Çetinkaya-Uzun score, it was 85.53 (SD=4.0, 5th-7th grade level, independently readable). The mean total DISCERN score was 54.55 (SD=5.75, which indicates good quality), and the mean GQS was 4.35 (SD=0.75, which suggests between good and excellent). No significant correlation was found between DISCERN and GQS scores (p=0.831).

Conclusion: The responses provided by the ChatGPT o1 model to patients' most frequently asked questions about colorectal cancer have medium-level readability and good-quality content. Therefore, it can be considered a helpful resource for patients seeking information.

Keywords: Artificial intelligence in oncology, cancer education, colorectal cancer

Introduction

Artificial intelligence (AI) language models such as ChatGPT (OpenAI Inc., California, United States) have become one of the most frequently used information sources by patients and their relatives since their introduction into daily use [1]. Numerous studies in the literature demonstrate that these models have sufficient medical knowledge, comparable to the level required to successfully pass medical licensing exams in various countries [2,3]. Based on these studies, it is believed that these models can provide appropriate answers to patients' questions. Studies prepared with this assumption have shown that healthcare providers indeed respond appropriately to patient inquiries [4].

The readability and quality of medical information obtained from the internet are among the biggest sources of concern. Therefore, many different analysis techniques have been developed for readability assessment. Methods such as DISCERN and the Global Quality Score (GQS) have been devised for evaluating content quality [5,6]. For Turkish publications, readability scores like Ateşman and Çetinkaya-Uzun are available for readability assessment and are frequently used in research [5]. Information can be obtained from various online sources such as videos, blogs, news sites, and forums. The comprehensibility and readability of this information, especially for elderly individuals and those with low literacy levels, raise serious concerns [7].

With the increasing daily use of AI language models and the advantages provided by newly developed ones, it is likely that patients will use AI language models like ChatGPT more frequently to access information. On 12/09/2024, OpenAI introduced the ChatGPT o1 model, which was designed to

Address for Correspondence: Efe Cem Erdat MD, Ankara University Faculty of Medicine, Department of Medical Oncology, Ankara, Türkiye
E-mail: cemerdat@gmail.com **ORCID ID:** orcid.org/0000-0002-1250-1297
Received: 24.10.2024 **Accepted:** 16.06.2025 **Epub:** 04.08.2025

Cite this article as: Erdat EC, Yalçıner M, Utkan M. Readability and quality analysis of ChatGPT o1's responses on colorectal cancer: a study of an AI language model. Acta Haematol Oncol Turc. [Epub Ahead of Print]

offer higher education, especially doctoral-level information. However, information regarding the responses of this model to patient questions is rarely found in the medical literature [8].

Methods

Data Collection

Our study was planned as a bibliographic study aimed at performing readability and quality analyses. The questions that patients searched for on the internet regarding colorectal cancer were obtained from the Google statistics (Google LLC, California, United States) and Yandex statistics (Yandex LLC, Moscow, Russia) search engines. Since there was no time limitation in Google statistics, information from 19/07/2019 to 18/07/2024 was collected. In Yandex statistics, due to a one-month limit on statistical data, information from 18/06/2024 to 18/07/2024 was obtained. After obtaining the statistics, the top 20 most searched topics were identified.

Once the topics were determined, questions were sequentially posed to the ChatGPT o1 model, and the obtained responses were saved as plain text files. To complete the readability analyses in the plain text files, punctuation and spelling errors were manually corrected. ChatGPT was not informed about the purpose of asking the questions, as the study aimed to evaluate the quality and readability of patients' questions.

Readability Analyses

Two different readability analysis methods specifically developed for the Turkish language were used.

The first analysis was conducted using the readability analysis developed by Ateşman [9] and published in 1997. The readability analysis developed by Ateşman [9] is an adaptation of the Flesch formula to Turkish which was originally developed for English. The formula is as follows: Readability score = 198.825 – 40.175 (x_1, average syllables per word) – 2.610 (x_2, average words per sentence). According to the formula developed by Ateşman [9], readability levels are determined as follows: 1-29: very difficult; 30-49: difficult; 50-69: moderately difficult; 70-89: easy; and 90-100: very easy.

The second analysis was conducted using the readability analysis developed by Çetinkaya and Uzun [10] and published in 2010. The readability analysis developed by Çetinkaya-Uzun is based on whitespace identification, and the formula is as follows: Readability score = 118.823 – (25.987 × average word length) – (0.971 × average sentence length). According to the formula developed by Çetinkaya-Uzun, readability levels are determined as follows: 0-34: insufficient reading level, corresponding to 10^{th}-12^{th} grade; 35-50: educational reading level, corresponding to 8^{th}-9^{th} grade; ≥51: independent reading level, corresponding to 5^{th}-7^{th} grade.

Simple code was written using Python 3.12 for readability analysis, and the analysis was performed on plain text files.

Quality Analyses

For quality analyses of the obtained materials, the DISCERN score and the GQS were used.

The DISCERN score was developed in English in 1998 and consists of 16 questions. Among these questions, 1-8 are about reliability, 9-15 are about treatment options, and question 16 is about overall quality. Each question is scored between 1 (poor) and 5 (good), and the total score is used for analysis. The recommended evaluation for the DISCERN score is as follows: 16-29: very poor; 30-40: poor; 41-51: fair; 52-63: good; 64-80: excellent.

The GQS is a simple scoring system ranging from 1 to 5. According to this score: 1: very poor; 2: poor; 3: fair; 4: good and 5: excellent.

Quality analyses were conducted by two different observers. Since there was complete agreement between them, the scores were assigned identically.

Statistical Analysis

Statistical analyses were conducted using GraphPad Prism 10 (GraphPad Inc., New Jersey, United States). For descriptive statistics, the mean and standard deviation (SD) were used. Pearson correlation analysis was employed to evaluate the relationships between scores; and one-way analysis of variance (ANOVA) was used to analyze different scores according to topics. A p value of less than 0.05 was considered statistically significant.

The observers' comments and the obtained texts were subjected to qualitative analysis techniques, with general thematic analyses also conducted. Qualitative data analysis was performed manually, identifying recurring words and themes. Graphs for the qualitative analysis were created using Python 3.12 and the "matplotlib" package.

Ethics Statement

Since the study was bibliographic in nature, ethical committee approval was not deemed necessary. The ChatGPT AI system was only used during the data collection phase, and it was not utilized in any analyses. The study was conducted in accordance with current and universal ethical standards.

Results

Readability Analysis

Twenty of the most frequently searched topics were obtained from the Google and Yandex search engines. When these topics were provided to the ChatGPT o1 model, the average number of words in the generated responses was calculated to be 654.9 (SD=221.62). According to Ateşman's [9] readability formula, the average readability score was 55.45 (SD=6.06), which was evaluated as moderately difficult to read. According to the Çetinkaya-Uzun readability formula, the average readability score was 85.53 (SD=4.0), and it was assessed as independently readable at the 5^{th}-7^{th} grade level. In the Pearson correlation analysis, R^2=0.395 was calculated and deemed statistically significant (p=0.003). The ranking of the obtained topics by frequency, word counts, Ateşman [9] readability scores, and Çetinkaya-Uzun readability scores is presented in Table 1 and Figure 1.

Quality Analysis

According to the DISCERN quality analysis, which consists of sixteen questions, Question 1, ("Is it relevant?") and Question 15 ("Does it provide support for shared decision-making?") received a score of 5 in all topics. Question 4 ("Are the sources of information used in compiling the publication clearly stated?") and Question 5 ("Is it clear when the published information is being used or reported?") received a score of 1 across all topics because no information was provided. The average score for the responses to the questions was 3.41 (SD=1.68). When all questions were evaluated together, significant score differences were observed with the use of the ANOVA test ($p<0.001$, $F=79.82$). The results of the scoring of DISCERN quality analysis questions are detailed in Figure 2.

Table 1. Determined topic headings, word count, and readability analysis

Topic heading	Word count	Ateşman readability	Çetinkaya-Uzun readability
Rectal cancer symptoms	161	91.03	56
Colon cancer	251	89.34	48
Colon cancer symptoms	387	87.12	62
Rectal cancer surgery	613	81.12	56
Rectal cancer treatment	805	76.85	42
What is rectal cancer?	621	85.03	58
Intestinal cancer symptoms	536	88.76	62
What causes rectal cancer?	559	90.45	57
What is colon cancer?	753	84.70	60
Blood from the anus	692	86.08	58
Colon and rectal cancer	1001	83.96	52
Experiences of those with rectal cancer	712	79.43	47
Stage 1 rectal cancer symptoms	544	88.86	65
Experiences of those with colon cancer	855	78.16	47
What causes colon cancer?	712	87.72	51
Colon cancer life expectancy	584	88.76	61
Stage 4 colon cancer	944	84.94	60
Rectal cancer life expectancy	976	84.52	58
Stage 3 colon cancer survival chance	798	86.30	52
What is a colon?	594	87.50	57
Average (SD)	**654.9 (SD=221.62)**	**85.53 (SD=4.00)**	**55.45 (SD=6.06)**

SD: Standard deviation

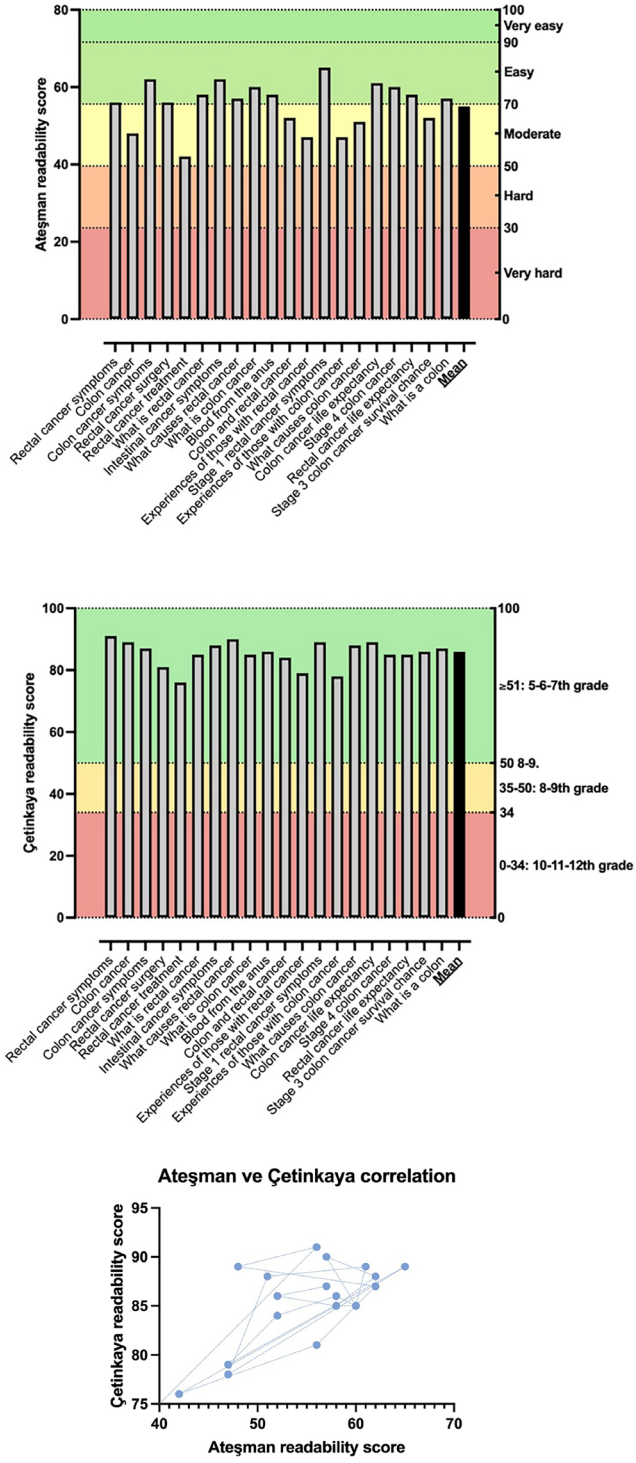

Figure 1. Ateşman and Çetinkaya-Uzun readability scores and their correlation

The total DISCERN score was calculated to have an average of 54.55 (SD=5.75), with the lowest score being 43 for the "colon cancer" topic and the highest score being 66 for the "blood from the anus" topic. The average total DISCERN score was classified as good. It was observed that 13 topics (65%) could be described as good, 6 topics (30%) as fair, and 1 topic (5%) as excellent. It was noted that none of the topics could be evaluated as poor or very poor according to the DISCERN analysis of the ChatGPT o1 model. The GQS score had an average of 4.35 (SD=0.75), indicating a rating between good and excellent. It was observed that 10 topics (50%) received a score of 5 and could be evaluated as excellent, 7 topics (35%) received a score of 4, evaluated as good, and 3 topics received a score of 3, evaluated as fair. In the correlation analysis, no significant correlation was observed between the DISCERN scores and GQS scores (R^2=0.002, p=0.831). The total DISCERN and GQS scores are presented in Table 2 and Figure 3.

A comparative analysis was not performed because the evaluators' scores were consistent.

Qualitative Analyses

In the qualitative analyses conducted, independent of the topic headings, the themes of the content were identified as cancer definitions, symptoms, risk factors, diagnostic methods, treatment options, life expectancy and prognosis, prevention and early diagnosis, and quality of life. Particularly, recurring content included definitions of colon and rectal cancer, risk factors, explanations of treatment methods, and recommendations for early diagnosis and screening.

The nine most frequently used words across all texts were

Table 2. Total DISCERN and GQS scores for determined topic headings

Topic heading	Total DISCERN score	GQS score
Rectal cancer symptoms	44	5
Colon cancer	43	5
Colon cancer symptoms	45	4
Rectal cancer surgery	57	5
Rectal cancer treatment	61	5
What is rectal cancer?	54	4
Intestinal cancer symptoms	51	4
What causes rectal cancer?	53	4
What is colon cancer?	58	4
Blood from the anus	66	5
Colon and rectal cancer	57	4
Experiences of those with rectal cancer	56	5
Stage 1 rectal cancer symptoms	51	5
Experiences of those with colon cancer	56	5
What causes colon cancer?	51	3
Colon cancer life expectancy	57	4
Stage 4 colon cancer	59	5
Rectal cancer life expectancy	58	3
Stage 3 colon cancer survival chance	56	3
What is a colon?	58	5
Average (SD)	**54.55 (SD=5.75)**	**4.35 (SD=0.75)**

SD: Standard deviation, GQS: Global Quality Score

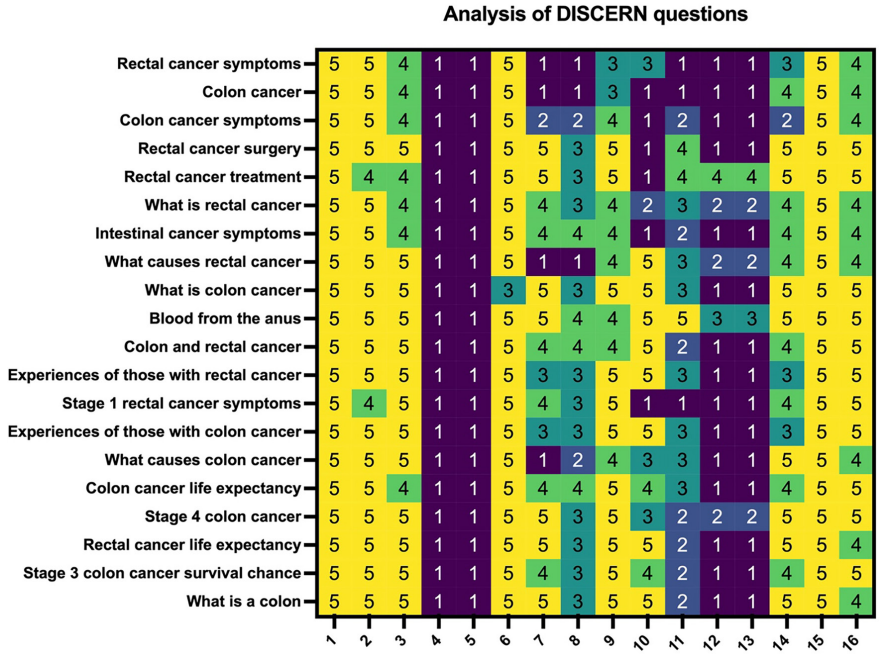

Figure 2. Analysis of DISCERN questions according to topic headings

observed to be “cancer” (215 occurrences), “colon” (189), “treatment” (142), “symptoms” (98), “stage” (87), “surgery” (76), “chemotherapy” (64), “life” (61), and “risk” (59). The frequency and cross-connections of the words used in the text are presented in Figure 4.

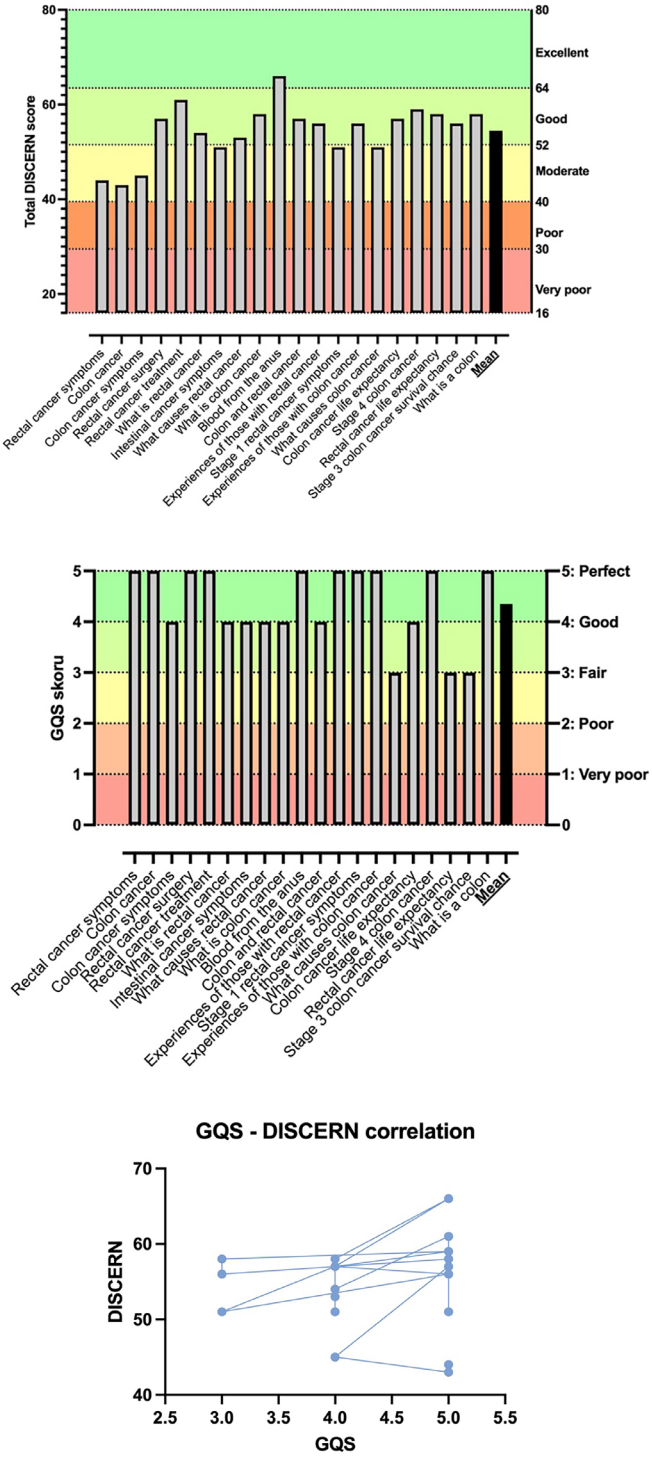

Figure 3. DISCERN and Global Quality Score (GQS) quality analysis and correlation

Discussion

In our study, the readability and quality levels of the responses provided by the ChatGPT o1 model to the most frequently searched patient questions related to colorectal cancer were examined. The results indicate that the responses from the ChatGPT o1 model possess a moderate level of readability and good quality.

Readability is critically important for patients to understand and apply health-related information. The Ateşman [9] and Çetinkaya and Uzun [10] readability formulas are reliable tools for determining the readability levels of Turkish texts. The readability scores obtained in our study demonstrate that the responses from the ChatGPT o1 model are generally understandable to the public. Specifically, according to the Çetinkaya-Uzun score, the texts are at a 5th-7th grade reading level, indicating that even individuals with low education levels can comprehend this information. This readability level is consistent with the internationally accepted 6th-grade readability standard for medical articles aimed at the public [11]. Additionally, it was observed that English terms that may slightly reduce comprehensibility were included in the responses generated by the ChatGPT o1 model. A limitation of the readability formulas is that they are solely based on words, syllables, and sentences. Therefore, the readability scores do not account for words originating from other languages.

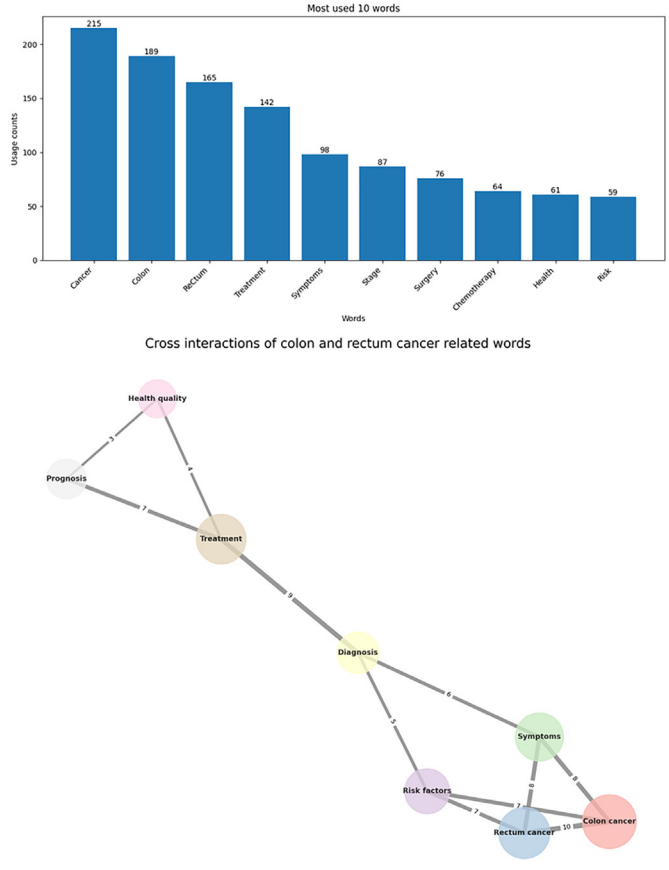

Figure 4. Word frequency and cross-connections graph in qualitative analysis

The DISCERN and GQS scores used in the quality assessment provide important information about the reliability and usability of health information materials. The average total DISCERN score is in the good quality range, and the GQS scores range between good and excellent. This suggest that the ChatGPT o1 model is capable of generating responses that meet patients' informational needs. However, the requirement for references in the DISCERN score and the inclusion of questions regarding the benefits and harms of all types of treatments, pose challenges. As these aspects cannot be adequately addressed within the generated texts based on the topic headings, the DISCERN score is insufficient for evaluating AI language models. Customized scoring systems appear to be necessary for the medical evaluation of texts generated by AI language models.

When the obtained scores are compared with other online sources, they can be considered to be of higher quality. It has been observed that approximately one-third of internet videos related to colorectal cancer and cancer screening are of poor quality in terms of information [12]. Additionally, publications report the inadequacy of online information sources concerning potential adverse events following rectal surgery [13]. Furthermore, information obtained from commercially operating websites carries a significant risk of bias [14].

Moreover, there is a risk of generating incorrect information, referred to as "hallucinations" in AI terminology [15]. These findings support the notion that AI language models could be a resource for accessing information in the health sector. However, due to the risk of hallucinations, caution is necessary.

Study Limitations

Our study has several limitations. Firstly, the research focused solely on the top 20 frequently searched topics related to colorectal cancer; therefore, the results may not be generalizable to all types of cancer or medical subjects. Additionally, readability and quality assessments were conducted using specific formulas and scales; the subjective nature of these methods may influence the results. Furthermore, the evaluations are based only on the performance of the ChatGPT o1 model within a specific time frame; future updates to the model and the emergence of more advanced models could alter these findings.

Conclusion

This study demonstrated that the responses provided by the ChatGPT o1 model to the most frequently asked patient questions regarding colorectal cancer have a moderate level of readability and good quality. The findings suggest that the model is a helpful resource for patients in accessing information.

Looking ahead, the implementation of AI in patient knowledge is poised to become even more transformative. Future advancements will likely enhance the accuracy and personalization of the information provided. AI models could integrate real-time updates from the latest medical research, ensuring that patients receive the most current and relevant information.

Moreover, the potential for AI to support patient education is immense. With the development of more sophisticated language models, AI could offer tailored educational content based on individual patient needs and learning styles. As AI continues to evolve, it holds the promise of empowering patients with the knowledge they need to make informed decisions about their health.

Ethics

Ethics Committee Approval-Informed Consent: Since the study was bibliographic in nature, ethical committee approval was not deemed necessary. The ChatGPT AI system was only used during the data collection phase, and it was not utilized in any analyses. The study was conducted in accordance with current and universal ethical standards.

Footnotes

Authorship Contributions

Concept: E.C.E., G.U., Design: G.U., Data Collection or Processing: E.C.E., M.Y., Analysis or Interpretation: E.C.E., M.Y., G.U., Literature Search: E.C.E., Writing: E.C.E.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Walker HL, Ghani S, Kuemmerli C, 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.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
3. Al-Shakarchi NJ, Haq IU. ChatGPT performance in the UK medical licensing assessment: how to train the next generation? Mayo Clin Proc Digit Health. 2023;1:309-310.
4. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol. 2024;21:353-359.
5. Kalyoncu MR, Memiş M. Consistency query and comparison of readability formulas created for Turkish. Journal of Mother Tongue Education. 2024;12:417-436.
6. Cakmak G. Evaluation of scientific quality of YouTube video content related to umbilical hernia. Cureus. 2021;13:e14675.
7. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res. 2022;24:e34790.
8. OpenAI Inc. (2024) Introducing OpenAI o1-preview. Last Accessed Date: 15.10.2024. Available from: https://openai.com/index/introducing-openai-o1-preview/
9. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi. 1997;58:71-74.
10. Çetinkaya G, Uzun L. Identifying and classifying the readability levels of the Turkish texts. Ankara Universitesi, Doctorate Thesis, 2010.

11. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp. 2021;8:2374373521998847.
12. Brar J, Ferdous M, Abedin T, Turin TC. Online information for colorectal cancer screening: a content analysis of YouTube videos. J Cancer Educ. 2021;36:826-831.
13. Brissette V, Alnaki A, Garfinkle R, et al. The quality, suitability, content and readability of online health-related information regarding sexual dysfunction after rectal cancer surgery. Colorectal Dis. 2021,23:376-383.
14. Li JZH, Kong T, Killow V, et al. Quality assessment of online resources for the most common cancers. J Cancer Educ. 2023;38:34-41.
15. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15:e35179.