ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.31609/jpmrs.2024-103532

Evaluation of Artificial Intelligence Supported Osteoarthritis Information Texts: Content Quality and Readability Analysis

Yapay Zekâ Destekli Osteoartrit Bilgilendirme Metinlerinin Değerlendirilmesi: İçerik Kalitesi ve Okunabilirlik Analizi

İlhan Celil ÖZBEK[a]

[a]University of Health Sciences Kocaeli Derince Training and Research Hospital, Department of Physical Medicine and Rehabilitation, Kocaeli, Türkiye

ABSTRACT Objective: This study aims to comprehensively evaluate the quality,readability,and understandability of artificial intelligence-supported texts related to osteoarthritis(OA). **Material and Methods:** The most frequently searched keywords related to osteoarthritis were determined through Google Trends. Additionally, frequently asked questions by patients about osteoarthritis were identified. These keywords and questions were entered into ChatGPT. The Ensuring Quality Information for Patients tool(EQIP) was used to assess the clarity of information and quality of writing.Flesch-Kincaid-readability-tests (Reading-Ease and Grade-Level) and Gunning-Fog-Index (GFI) were used to assess the readability of the texts.The reliability and usefulness of the texts were assessed were used the reliability and usefulness scale. **Results:** The average scores were: EQIP 62.01±6.61, FKRE 31.85±12.44, FKGL 13.26±2.12,GFI 14.52±2.41, reliability 5.10±1.02,and usefulness 4.89±0.76. Our study concludes that ChatGPT's responses on osteoarthritis are generally of "good-quality with minor-issues". Additionally, it was determined that the texts produced were of complexity that they would require approximately 13 years of education. When the EQIP score obtained from texts created using keywords was compared with the EQIP score obtained from texts created using questions,a statistically significant difference was observed ($p<0.001$). However, when examined in terms of FKRE, FKGL, GFI, Reliability-Scale and Usefulness-Scale scores between the two groups,no statistically significant difference was found. (respectively, p=0.063, p=0.059, p=0.194, p=0,466, p=0,499). **Conclusion:** This-study reveals that ChatGPT's texts on OA have certain deficiencies in quality and readability.In conclusion,it emphasizes that online resources and AI tools play an important role in information provision in the field of healthcare,but quality and readability control should be ensured.In addition to ensuring patients have access to accurate,reliable and understandable information,this can help them make more informed and effective health decisions by increasing their health literacy.

Keywords: ChatGPT; quality assessment; readability; osteoarthritis

ÖZET Amaç: Bu çalışmanın amacı,osteoartrit ile ilgili yapay zeka destekli oluşturulan metinlerin içeriğinin kalitesini,okunabilirliğini ve anlaşılabilirliğini kapsamlı bir şekilde değerlendirmektir. **Gereç ve Yöntemler:** Google Trends üzerinden osteoartrit ile ilgili en sık aranan anahtar kelimeler belirlendi. Belirlenen anahtar kelimelerle birlikte,osteoartrit hakkında hasta tarafından sıkça sorulan sorular seçildi. Belirlenen anahtar kelimeler ve sorular sırayla ChatGPT'ye girildi. Belirlenen anahtar kelimeler ve sorular ChatGPT'ye aktarıldı. Bilginin netliği ve yazım kalitesini değerlendirmek için Hastalar için Kaliteli Bilgi Sağlama aracı (EQIP) kullanıldı. Metinlerin okunabilirliğini değerlendirmek için Flesch–Kincaid okunabilirlik testleri (Okuma Kolaylığı ve Sınıf Düzeyi) ve Gunning Fog İndeksi (GFI) kullanıldı. Metinlerin güvenilirliği ve yararlılığını, güvenilirlik ve yararlılık ölçeği kullanılarak değerlendirildi. **Bulgular:** Metinlerin ortalama EQIP skoru 62,01±6,61'di. Flesch–Kincaid Okuma Kolaylığı (FKRE) ortalama skoru ise 31,85±12,44'tü. Flesch–Kincaid Sınıf Düzeyi (FKGL) için ortalama skor 13,26±2,12'ydi. GFI skoru ortalaması ise 14,52±2,41'di. Metinlerin ortalama Güvenilirlik puanı 5.10±1.02'di. Metinlerin ortalama Yararlılık puanı 4,89±0,76' dı. Çalışmamız, ChatGPT'nin osteoartrit konusundaki yanıtlarının genel olarak "küçük sorunlarla birlikte iyi kaliteli" olduğu sonucuna varmaktadır. Ayrıca, üretilen metinlerin yaklaşık 13 yıl eğitim gerektirecek karmaşıklıkta olduğu belirlendi. Anahtar kelimeler kullanılarak oluşturulan metinlerden elde edilen EQIP skoru ile sorular kullanılarak oluşturulan metinlerden elde edilen EQIP skoru karşılaştırıldığında,istatistiksel olarak anlamlı bir farklılık gözlemlenmiştir ($p<0.001$). Ancak, iki grup arasında FKRE, FKGL, GFI, Güvenilirlik ölçeği ve Yararlılık ölçeği skorları açısından incelendiğinde, istatistiksel olarak anlamlı bir farklılık bulunmamıştır. (sırasıyla, p=0.063, p=0.059, p=0.194, p=0,466, p=0,499). **Sonuç:** Bu çalışma, ChatGPT'nin osteoartrit hakkındaki metinlerinin kalite ve okunabilirlik konusunda belirli eksikliklerin bulunduğunu ortaya koymaktadır. Sonuç olarak, çevrimiçi kaynakların ve yapay zeka araçlarının sağlık alanında bilgi sunumunda önemli bir rol oynadığını, ancak kalite ve okunabilirlik kontrolünün sağlanması gerektiğini vurgulamaktadır. Bu, hastaların doğru, güvenilir ve anlaşılır bilgilere erişimini sağlamanın yanı sıra,sağlık okuryazarlığını artırarak daha bilinçli ve etkin sağlık kararları alabilmelerine yardımcı olabilir.

Anahtar Kelimeler: ChatGPT; kalite değerlendirmesi; okunabilirlik; osteoartrit

Correspondence: İlhan Celil ÖZBEK
University of Health Sciences Kocaeli Derince Training and Research Hospital, Department of Physical Medicine and Rehabilitation, Kocaeli, Türkiye
E-mail: ilhanozbek7@gmail.com

Peer review under responsibility of Journal of Physical Medicine and Rehabilitation Science.

Received: 19 Apr 2024 ***Received in revised form:*** 07 Nov 2024 ***Accepted:*** 13 Nov 2024 ***Available online*:** 27 Nov 2024

Osteoarthritis (OA) is the most common type of arthritis, typically affecting two major joints such as the knee and hip.[1,2] This disease causes various anatomical and physiological changes, including cartilage degradation, bone remodeling, and osteophyte formation.[3] It is known to affect over 500 million people worldwide, with one in every three individuals over the age of 65 struggling with OA.[1,4] With the rapidly aging population and changing lifestyle habits, the prevalence of OA has significantly increased and is projected to continue rising in the future. Considering the growing burden and impact of OA, more efforts are needed to provide effective and safe treatments to those dealing with this condition.[1]

Artificial intelligence (AI) is used to denote the development of algorithms designed to perform tasks typically associated with intelligent behavior, often performed by humans. These tasks encompass areas such as natural language understanding, image recognition, decision-making, problem-solving, and learning from experiences.[5] AI finds applications in various fields within the healthcare sector, including medical imaging, diagnosis, decision support systems, drug discovery and development, patient monitoring, robot-assisted surgery, and virtual assistants.[6]

AI-powered virtual assistants and chatbots can address patients' needs, answer their questions, provide basic health information, and assist in scheduling appointments.[7] ChatGPT is an AI language model developed by the US-based company OpenAI, used for natural language processing and text generation. Trained on extensive text datasets, this system is designed to produce appropriate and consistent human-like responses to user inputs. Thanks to its ability to provide fast and detailed responses along with its accessibility, ChatGPT reached 100 million users just two months after its launch.[8]

Health literacy generally decreases with age. Since OA is an age-related condition, providing accessible, understandable, and reliable information to patients is even more crucial. Well-structured and reliable information can help patients understand the details of the disease, treatment options, and preventive measures.[9,10] Numerous studies investigating the quality and readability of health information related to medical conditions are available in the literature. However, the number of studies evaluating health information generated by ChatGPT, particularly regarding OA, is quite limited.

The aim of this study is to comprehensively assess the quality and readability of AI-generated texts related to OA.

MATERIAL AND METHODS

The study was conducted at our University's School of Medicine Hospital between April 13 and April 19, 2024. No human participants or animals were involved in this study; therefore, ethical approval was not required. Similar studies in the literature have followed the same approach.[11]

Google Trends (Google, USA) (https://trends.google.com) was used to identify the most frequently searched keywords related to OA. Before initiating the searches, all browser-related data were completely cleared in a manner that would not affect the results. Searches were conducted separately for "Osteoarthritis," "Knee osteoarthritis," and "Hip osteoarthritis." The search criteria included data from all regions worldwide and across all categories, spanning the period from 2004 to the present. The most relevant words were selected in the relevant queries section in the results. The top twenty-five keywords from each search and a total of 75 keywords were recorded. The exclusion criteria of the study comprised a total of 54 keywords, consisting of 49 repetitive terms and 5 irrelevant terms, and were thus excluded from the analysis. In addition to the keywords identified, 7 questions commonly asked by patients about hip and knee OA were determined (Table 1).[12]

Initially, a dedicated account was created for this study. The selected 21 keywords and 7 questions were sequentially entered into the chat interface of the ChatGPT AI chatbot version April 13 (Table 2). Each keyword and question were processed in separate chat pages to minimize the potential impact of previous queries and responses. The resulting answers were methodically documented, focusing particularly on quality, comprehensiveness, and readability for subsequent analyses.

TABLE 1: Most commonly asked questions about OA.

What can I do myself to decrease OA symptoms and to prevent the OA from getting worse?
What is the natural course of OA?
What are the newest treatment options for OA?
Is there any medication that can either slow down or stop OA?
What can or can I not do in terms of exercise and physical activity for OA?
I'm young and I have OA. What changes should I make to my life and what should or shouldn't I do anymore?
Can exercise or being physically active be harmful to my joints?

OA: Osteoarthritis.

TABLE 2: Most frequently searched keywords related to osteoarthritis.

Osteoarthritis	Osteoarthritis causes	Osteoarthritis symptoms
Knee osteoarthritis	Osteoarthritis exercises	Knee osteoarthritis symptoms
Hip osteoarthritis	Knee osteoarthritis exercises	Hip osteoarthritis symptoms
Arthritis	Hip osteoarthritis exercises	Knee replacement
Osteoarthritis treatment	Osteoarthritis pain	Hip replacement
Hip osteoarthritis treatment	Knee pain	Knee joint
Knee osteoarthritis treatment	Hip pain	Hip joint

To assess the clarity and writing quality of the obtained 29 texts, the tool for Ensuring Quality Information for Patients (EQIP) used. EQIP comprises 20 items. Each item is evaluated with responses of "yes," "partially," "no," or "not applicable (N/A)".[13]

Scoring of the EQIP tool was done as follows: "yes" answers received 1 point, "partially" answers received 0.5 points and "no" answers received 0 points. Items marked as "not applicable" were subtracted from the total number of items. The total score obtained was divided by the number of valid items and calculated as a percentage. EQIP scores were categorized into different groups according to the ranges recommended in the EQIP development publication: sources scoring between 76% and 100% were classified as "well-written and high quality", those scoring between 51% and 75% were classified as "good quality with minor issues", those scoring between 26% and 50% were classified as "significant quality issues", and those scoring between 0% and 25% were also classified as "significant quality issues".[14]

Each text was independently assessed by two physical medicine and rehabilitation specialists with at least 5 years of experience (EO and CU) in different settings to minimize bias. Differing scores due to the subjective nature of some questions were resolved by the author (ICO) to reach a consensus. After resolving the inconsistent scores, the two EQIP scores calculated for each source were averaged.

To assess the readability of the texts, the Flesch-Kincaid readability tests (Readability Ease and Grade Level) and the Gunning Fog Index (GFI) were used.

The Flesch-Kincaid Readability Ease (FKRE) Score is calculated using the formula: 206.835-(1.015 x average sentence length)-(84.6 x average syllables per word). The higher the score on the test, the more readable the content is. A score below 30 indicates a reading level comparable to that of university graduates.[15]

The Flesch-Kincaid Grade Level (FKGL) Score is calculated using the formula: 0.39x (Total words/Total sentences) +11.8x(Total syllables/Total words)-15.59. The result indicates the educational level of the audience the text is aimed at. For example, a result of 10 and above suggests the text is aimed at a high school level audience.[15]

The GFI is an assessment based on sentence length and the complexity of words. GFI is calculated using the formula: [(number of words/number of sentences)+("number of words with three or more syllablesx100"/"number of words")]x0.4. According to the formula, shorter sentences indicate better readability. A score above 12 indicates a difficult text to read.[15]

The reliability and usefulness of the texts were assessed using the reliability and usefulness scale developed by Uz and Umay.[16] In the reliability scale, it is evaluated whether the answers can be verified from scientific sources and if they contain any incomplete or incorrect information. The scale ranges from a minimum score of 1 to a maximum of 7, with higher scores indicating higher reliability.[16]

In the usefulness scale, the understandable of the answers and whether the provided information is beneficial for patients are evaluated. Similarly, the scale ranges from a minimum score of 1 to a maximum of

7, with higher scores indicating greater usefulness.[16]

Each text was independently assessed by two physical medicine and rehabilitation specialists with at least 5 years of experience (ICO and CU) in different settings to minimize bias. The two reliability and usefulness scores calculated for each source were averaged.

This research was conducted according to the principles set out in the Declaration of Helsinki.

The study data were analyzed using the SPSS 27.0 software (IBM, USA) package. Descriptive statistics were presented as mean±standard deviation for variables with a normal distribution, and as median (minimum-maximum) for variables with a non-normal distribution. The normality of variables was assessed visually using histograms and probability plots, and analytically using the Kolmogorov-Smirnov test. Differences between groups in terms of continuous variables were investigated using the Mann-Whitney U test for independent groups. Spearman correlation test was used for correlation analysis of numerical data. Results were considered statistically significant for $p<0.05$.

RESULTS

The mean, standard deviation, minimum, and maximum values of EQIP, FKRE, FKGL, GFI, Reliability Scale and Usefulness Scale scores are presented in Table 3.

The EQIP scores, which assess the writing quality of the texts, ranged from 45.45% to 73.73%, with a mean value of 62.01%±6.61, indicating an overall moderate writing quality. For readability measures, the FKRE scores varied between 6.80 and 66.90, with an average score of 31.85, reflecting that most texts were moderately challenging to read. The FKGL scores ranged from 7.10 to 18.10, averaging 13.26±2.12, suggesting a readability level suited for audiences with a high school to early college reading level. The GFI scores ranged from 8.39 to 20.23, with a mean score of 14.52±2.41, indicating that the texts varied widely in complexity, from accessible to difficult.

The Reliability Scale, which evaluates the accuracy and verifiability of the information, yielded scores between 4 and 7, with a mean of 5.10, indicating generally reliable information. The Usefulness Scale, assessing the understandability and helpfulness for patients, scored between 4 and 7, with a mean score of 4.89, suggesting moderate usefulness across the texts.

Table 4 includes the average, standard deviation, minimum, and maximum values of EQIP, FKRE, FKGL, GFI, Reliability Scale and Usefulness Scale scores for texts generated with keywords and texts generated using questions.

When comparing texts generated with keywords against those generated in response to patient questions, a statistically significant difference was observed in EQIP scores ($p<0.001$). Texts generated with questions had significantly higher EQIP scores (mean difference of approximately 10%), indicating better writing quality than those generated with keywords. However, no statistically significant differences were found between the two groups for FKRE, FKGL, GFI, Reliability Scale, and Usefulness Scale scores (p-values were 0.063, 0.059, 0.194, 0.466, and 0.499, respectively).

TABLE 3: Statistics of EQIP, FKRE, FKGL, GFI, Reliability Scale and Usefulness Scale scores.

	Minimum	Maximum	Mean	Standard deviation
EQIP (%)	45.45	73.73	62.01	6.61
FKRE	6.80	66.90	31.85	12.44
FKGL	7.10	18.10	13.26	2.12
GFI	8.39	20.23	14.52	2.41
Reliability Scale	4.00	7.00	5.10	1.02
Usefulness Scale	4.00	6.00	4.89	0.76

EQIP: Ensuring Quality Information for Patients score; FKRE: The Flesch-Kincaid Reading Ease score; FKGL: The Flesch-Kincaid Grade Level score; GFI: Gunning Fog Index score.

TABLE 4: Comparison of EQIP, FKRE, FKGL, GFI, Reliability Scale and Usefulness Scale scores between texts generated with keywords and texts generated using questions.

	Keywords Group (n=21) $\bar{X}$±SD Minimum-Maximum	Questions Group (n=7) $\bar{X}$±SD Minimum-Maximum	The difference between groups (p)
EQIP (%)	59.39±5.10 45.45-66.66	69.84±3.85 63.63-73.33	<0.001
FKRE	34.50±12.01 15.90-66.90	23.88±10.83 6.80-37.40	0.063
FKGL	12.76±1.86 7.10-15.70	14.75±2.29 12.20-18.10	0.059
GFI	14.28±2.66 8.39-20.23	15.24±1.36 12.63-16.86	0.194
Reliability Scale	5.04±1.10 4-7	5.28±0.75 4-6	0.466
Usefulness Scale	4.85±0.82 4-6	5.00±0.57 4-6	0.499

EQIP: Ensuring Quality Information for Patients score; FKRE: The Flesch-Kincaid Reading Ease score; FKGL: The Flesch-Kincaid Grade Level score; GFI: Gunning Fog Index score; SD: Standard deviation.

Table 5 and Table 6 provide detailed scores of the EQIP, FKRE, FKGL, GFI, Reliability Scale, and Usefulness Scale for the keywords and responses generated to the questions.

In the correlation analysis, no statistically significant correlation was found between EQIP score and FKGL, FKRE, GFI, Reliability Scale and Usefulness Scale scores. However, a negatively high

TABLE 5: EQIP, FKRE, FKGL, GFI, Reliability and Usefulness Scale scores of texts created with keywords.

	EQIP	FKRE	FKGL	GFI	Reliability Scale	Usefulness Scale
Osteoarthritis	54.54	45.40	11.30	14.88	5.00	5.00
Knee osteoarthritis	58.33	27.90	13.80	13.92	6.00	6.00
Hip osteoarthritis	56.25	28.20	13.70	13.98	6.00	6.00
Arthritis	54.54	26.70	14.30	18.63	4.00	4.00
Osteoarthritis treatment	53.12	21.50	14.20	16.35	7.00	6.00
Hip osteoarthritis treatment	56.25	15.90	14.30	16.23	7.00	6.00
Knee osteoarthritis treatment	56.25	29.80	13.10	14.79	7.00	6.00
Osteoarthritis causes	63.63	29.40	13.20	14.49	6.00	6.00
Osteoarthritis exercises	63.33	28.50	13.60	13.28	4.00	4.00
Knee osteoarthritis exercises	63.33	28.40	13.60	13.40	4.50	4.50
Hip osteoarthritis exercises	63.33	66.90	7.10	8.39	4.00	5.00
Osteoarthritis pain	56.25	20.20	14.70	15.69	4.00	4.00
Knee pain	62.50	37.10	12.40	14.63	5.00	4.50
Hip pain	63.63	44.70	11.50	12.97	4.00	4.00
Osteoarthritis symptoms	63.63	46.70	10.70	10.15	4.00	4.50
Knee osteoarthritis symptoms	63.63	44.40	11.60	11.91	4.50	4.50
Hip osteoarthritis symptoms	63.63	45.30	11.30	11.01	5.00	5.00
Knee replacement	66.66	33.60	13.70	15.89	4.00	4.00
Hip replacement	60.00	23.00	15.70	20.23	4.00	4.00
Knee joint	59.09	45.60	11.20	13.60	6.00	5.00
Hip joint	45.45	35.50	13.00	15.61	5.00	4.00

EQIP: Ensuring Quality Information for Patients score; FKRE: The Flesch-Kincaid Reading Ease score; FKGL: The Flesch-Kincaid Grade Level score; GFI: Gunning Fog Index score.

TABLE 6: EQIP, FKRE, FKGL, GFI, Reliability and Usefulness Scale scores of the texts created using the questions.

	EQIP	FKRE	FKGL	GFI	Reliability Scale	Usefulness Scale
What can I do myself to decrease OA symptoms and to prevent the OA from getting worse?	71.87	37.40	12.20	12.63	6.00	6.00
What is the natural course of OA?	63.63	27.90	13.80	15.22	5.00	5.00
What are the newest treatment options for OA?	73.33	6.80	17.80	16.47	4.00	4.00
Is there any medication that can either slow down or stop OA?	65.62	11.30	18.10	16.86	5.00	5.00

EQIP: Ensuring Quality Information for Patients score; FKRE: The Flesch-Kincaid Reading Ease score; FKGL: The Flesch-Kincaid Grade Level score; GFI: Gunning Fog Index score; OA: Osteoarthritis.

TABLE 7: Correlation analysis data.

	EQIP	FKRE	FKGL	GFI	Reliability Scale	Usefulness Scale
EQIP	1	0.016	0.027	-0.165	-0.154	-0.064
	-	0.937	0.892	0.403	0.433	0.748
FKRE	0.016	1	**-0.979**	**-0.782**	-0.101	-0.007
	0.937	-	**<0.001**	**<0.001**	0.610	0.971
FKGL	0.027	**-0.979**	1	**0.822**	-0.003	-0.087
	0.892	**<0.001**	-	**<0.001**	0.988	0.662
GFI	-0.165	**-0.782**	**0.822**	1	0.029	-0.181
	0.403	**<0.001**	**<0.001**	-	0.884	0.356
Reliability Scale	-0.154	-0.101	-0.003	0.029	1	**0.868**
	0.433	0.610	0.988	0.884	-	**<0.001**
Usefulness Scale	-0.064	-0.007	-0.087	-0.181	**0.868**	1
	0.748	0.971	0.662	0.356	**<0.001**	-

r: 0.01-0.29 indicates a low level of correlation, r: 0.30-0.70 indicates a moderate level of correlation, r: 0.71-0.99 indicates a high level of correlation, p<0.05. Spearman correlation test; EQIP: Ensuring Quality Information for Patients score; FKRE: The Flesch-Kincaid Reading Ease score; FKGL: The Flesch-Kincaid Grade Level score; GFI: Gunning Fog Index score.

level statistically significant relationship was found between GFI score and FKRE score (p<0.001, r=-0.782). Additionally, a positively high level statistically significant relationship was detected between GFI score and FKGL score (p<0.001, r=0.822). A positively high level statistically significant relationship was detected between the scores of the Reliability Scale and the Usefulness Scale (p<0.001, r=0.868) (Table 7).

DISCUSSION

Our study concludes that ChatGPT's responses on the topic of OA are generally of "good quality with minor issues." It was determined that the average FKRE score is 31, indicating that the generated texts are of complexity requiring approximately 13 years of education. According to the reliability and usefulness scale, the responses were assessed as reliable and moderately useful. This is the first study evaluating the quality and readability of responses to the most frequently used keywords about OA and questions commonly asked by patients.

Accessible, accurate and easily understandable information is crucial in supporting individuals struggling with OA. Good quality and simple texts help patients to understand the complexity of the disease, available treatment options and preventive measures.

Barrow et al. conducted a study examining the readability and reliability of information available on OA on the internet. Their study revealed significant differences in quality among the evaluated websites.[17] Similarly, Anderson et al. emphasized the poor quality of online patient information sources related to OA.[18]

Chapman et al., in their research, noted low scores in readability for the information presented on web pages.[9] They emphasized that this situation implies that many individuals may not be able to read, understand, or effectively utilize the information.

Erden et al., in their study, stated that ChatGPT provides easily accessible information about osteoporosis, but they also noted shortcomings in terms of quality and readability.[11]

According to the results of our study, it was determined that the texts generated by ChatGPT exceeded standards for both quality and readability aimed at patients. However, EQIP evaluations showed that all examined texts were quite successful in aspects such as progressing logically, having a smooth layout, and addressing respectfully and personally. In some evaluation criteria, all analysed texts scored zero points. We believe that even small improvements at this point can move texts from the "good quality" category to the "well-written and high quality" category. We would like to emphasise that the main problem evident here is the lack of readability of the texts. To solve this problem, the importance of evaluating the quality of texts produced especially in the field of health with indices such as EQIP and readability indices such as FKRE, FKRL, GFI should be emphasized by teaching AI. In order to make the necessary arrangements, the database needs to be improved and audited. After these improvements, people, especially those who are not knowledgeable in the field of technology and health, will be able to gain a deeper understanding of their diseases and treatment processes.

In our study, the highest reliability and usefulness scores according to the evaluation scales were attributed to headings containing treatment inquiries. The information provided covered treatment options mentioned in the OA treatment guidelines of institutions like the American College of Rheumatology and the European League Against Rheumatism, including medications, physical therapy, lifestyle changes, injections, surgery, and alternative therapies.[19,20]

Even the lowest scores in terms of usefulness and reliability were determined to be beneficial and reliable for patients. However, some information gaps were present in the texts, and a more comprehensible language is necessary for patient education. These findings indicate that, while ChatGPT is generally a reliable and useful source for obtaining information about OA, there are areas that need improvement.

For instance, when we examined the response to the question "What can or can I not do in terms of exercise and physical activity for osteoarthritis?", we found both strengths and weaknesses. The strengths include providing a clear guide on which exercises can and cannot be done, making it easy for the reader to obtain information. Additionally, the response offers practical and applicable suggestions on how to perform the exercises and emphasizes the importance of safe exercise practices. Highlighting the necessity of consulting a specialist is also considered an important detail of the text.

The weaknesses of the text include a lack of detailed explanations regarding the types of exercises. More examples and descriptions could make the text more informative. Adding more scientific explanations about why exercises are beneficial and how they work could also enhance the text. Sometimes, the text uses complex and lengthy sentences; using simpler and more understandable language could increase accessibility.

Our study revealed that the texts classified under the heading "questions" have higher quality content with a statistically significant difference compared to the texts classified under the heading "keywords". However, although texts classified under the "keywords" category did not statistically differ from those under the "questions" category, they were evidently more readable. According to the reliability and usefulness scale, no statistical difference was observed between the two groups.

We clearly observed in our study that asking questions related to the topic rather than using simple keywords significantly influenced the quality and readability of the generated texts by ChatGPT. ChatGPT produces human-like texts in response and the quality and readability of responses may vary depending on the prompts. Asking questions related to the topic can help ChatGPT develop an understanding of the health literacy of the individual in front of it and produce responses tailored to their level. Responses to specific questions aimed at helping patients understand the details of their illnesses, treatment options, and preventive measures may involve more complex medical terms, which can affect

readability levels. Considering the overall readability challenges and high level of educational requirements of the generated texts, it is important to implement previously suggested strategies to improve their readability and quality.

With recent technological advancements, accessing information has become easier than ever before. Especially in health matters, everyone, including patients with OA and their caregivers, can easily gather information about relevant illnesses through online resources and recently popular chat tools like ChatGPT.[21] However, researches indicates significant quality and readability deficiencies in these online resources.[22-26]According to the results of our study, it is evident that ChatGPT also needs improvement in terms of quality and readability. In this context, individuals, particularly patients and their caregivers, may suffer due to access to misinformation in their quest for medical guidance.[27] Therefore, ensuring the accuracy, quality, and readability of information is of utmost importance. When these checks are in place, patients and their caregivers can access correct and reliable information. Therefore, special emphasis should be placed on quality and readability control in the information presentation process of online resources and tools like ChatGPT. This would be a step towards enhancing health literacy while ensuring patient safety. When these conditions are met, patients can become more aware and take an active role in understanding the importance of disease acceptance, treatment, and disease management.[28]

On the other hand, patients' complex medical conditions, varied medical and socio-cultural backgrounds, and symptoms should be addressed with a personalized assessment by medical professionals.[29] This ensures the creation of the most appropriate diagnosis and treatment plan. At this point, online resources and AI tools like ChatGPT cannot replace the role of healthcare professionals.[30] The uniqueness and critical importance of the physician-patient relationship should always be emphasized.

Although the number of keywords and related questions evaluated in our study is approximately similar to other studies of same nature, it may still create limitations in making generalizations. This can be considered a limitation of our study. Another limitation of this study is that only terms related to knee and hip OA were used as keywords. Therefore, it is not possible to comment on other types of OA. In future research, the inclusion of different types of OA may expand the scope of the study and increase the generalizability of the results.

CONCLUSION

Our study demonstrates that ChatGPT's responses regarding OA are generally of good quality, but they exhibit shortcomings in readability and some quality criteria. With an average FKRE score of 31, these texts are comprehensible at approximately a 13-year education level. According to reliability and utility evaluations, the responses are deemed reliable and moderately useful. Specifically, topics involving treatment inquiries received high reliability and utility scores, yet there is a need for clearer language and addressing information gaps.

It is emphasized that online resources and AI tools play a significant role in providing health information. However, ensuring quality and readability control is crucial. Continuous updates and improvements to ChatGPT and similar AI tools can enhance their potential to provide more effective and accessible health information.

Acknowledgements

I would like to thank Dr. Ceydanur Uçar and Dr. Emir Onağ for their assistance in evaluating the EQIP scores of the texts.

REFERENCES

1. Hawker GA, King LK. The burden of osteoarthritis in older adults. Clin Geriatr Med. 2022;38:181-92. PMID: 35410675.
2. Katz JN, Arant KR, Loeser RF. Diagnosis and treatment of hip and knee osteoarthritis: a review. JAMA. 2021;325:568-78. PMID: 33560326; PMCID: PMC8225295.
3. Allen KD, Thoma LM, Golightly YM. Epidemiology of osteoarthritis. Osteoarthritis Cartilage. 2022;30:184-95. PMID: 34534661; PMCID: PMC10735233.
4. Hunter DJ, March L, Chew M. Osteoarthritis in 2020 and beyond: a Lancet Commission. Lancet. 2020;396:1711-2. PMID: 33159851.
5. van Hartskamp M, Consoli S, Verhaegh W, et al. Artificial intelligence in clinical health care applications: viewpoint. Interact J Med Res. 2019;8:e12100. PMID: 30950806; PMCID: PMC6473209.
6. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69S:S36-S40. PMID: 28126242.
7. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell. 2023;6:1166014. PMID: 37091303; PMCID: PMC10113434.
8. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. 2024;245:108013. PMID: 38262126.
9. Chapman L, Brooks C, Lawson J, et al. Accessibility of online self-management support websites for people with osteoarthritis: a text content analysis. Chronic Illn. 2019;15:27-40. PMID: 29254372.
10. Varady NH, Dee EC, Katz JN. International assessment on quality and content of internet information on osteoarthritis. Osteoarthritis Cartilage. 2018;26:1017-26. PMID: 29758353.
11. Erden Y, Temel MH, Bağcıer F. Artificial intelligence insights into osteoporosis: assessing ChatGPT's information quality and readability. Arch Osteoporos. 2024;19:17. PMID: 38499716.
12. Claassen AAOM, Kremers-van de Hei KCALC, van den Hoogen FHJ, et al. Most important frequently asked questions from patients with hip or knee osteoarthritis: a best-worst scaling exercise. Arthritis Care Res (Hoboken). 2019;71:885-92. PMID: 30055092.
13. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7:165-75. PMID: 15117391; PMCID: PMC5060233.
14. Ladhar S, Koshman SL, Yang F, et al. Evaluation of online written medication educational resources for people living with heart failure. CJC Open. 2022;4:858-65. PMID: 36254325; PMCID: PMC9568683.
15. Benzer A. [A step toward the formula of readability based on artificial intelligence]. Research and Experience Journal. 2020;5:47-82. https://dergipark.org.tr/tr/download/article-file/1162032
16. Uz C, Umay E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis. 2023;26:1343-9. PMID: 37218530.
17. Barrow A, Palmer S, Thomas S, et al. Quality of web-based information for osteoarthritis: a cross-sectional study. Physiotherapy. 2018;104:318-26. PMID: 30030036.
18. Anderson KJ, Walker RJ, Lynch JM, et al. A qualitative evaluation of internet information on hip and knee osteoarthritis. Ann R Coll Surg Engl. 2023;105:729-33. PMID: 37489520; PMCID: PMC10618034.
19. Kolasinski SL, Neogi T, Hochberg MC, et al. 2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee. Arthritis Care Res (Hoboken). 2020;72:149-62. Erratum in: Arthritis Care Res (Hoboken). 2021;73:764. PMID: 31908149; PMCID: PMC11488261.
20. Moseng T, Vliet Vlieland TPM, Battista S, et al. EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis: 2023 update. Ann Rheum Dis. 2024;83:730-40. PMID: 38212040; PMCID: PMC11103326.
21. Gravina AG, Pellegrino R, Cipullo M, et al. May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis. World J Gastroenterol. 2024;30:17-33. PMID: 38293321; PMCID: PMC10823903.
22. Temel MH, Batıbay S, Bağcıer F. Quality and readability of online information on cerebral palsy. Journal of Consumer Health on the Internet. 2023;27:266-81. https://doi.org/10.1080/15398285.2023.2235531
23. Hong SW, Kang JH, Park JH, et al. Quality and readability of online information on hand osteoarthritis. Health Informatics J. 2023;29:14604582231169297. PMID: 36995242.
24. Ozduran E, Hanci V. Evaluating the readability, quality, and reliability of online information on Sjogren's syndrome. Indian Journal of Rheumatology. 2023;18:16-25. https://avesis.deu.edu.tr/yayin/df966002-0da6-4c74-bc5f-1748be89dd72/evaluating-the-readability-quality-and-reliability-of-online-information-on-sjogrens-syndrome
25. Kaya E, Görmez S. Quality and readability of online information on plantar fasciitis and calcaneal spur. Rheumatol Int. 2022;42:1965-72. PMID: 35763090.
26. Hanci V, Bıyıkoğlu BO, Bıyıkoğlu AS. How readable the online patient education materials of intensive and critical care societies: assessment of the readability. Journal of Critical Care. 2024;81:154713. https://doi.org/10.1016/j.jcrc.2024.154713
27. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel). 2024;14:109. PMID: 38201418; PMCID: PMC10802884.
28. Coşkun AB, Elmaoğlu E, Buran C, et al. Integration of Chatgpt and E-health literacy: opportunities, challenges, and a look towards the future. Journal of Health Reports and Technology. 2024;10:e139748. https://doi.org/10.5812/jhrt-139748
29. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg. 2024;181:e1138-e44. PMID: 38000671.
30. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023;3:100105. https://www.sciencedirect.com/science/article/pii/S2772485923000224